Universidade do Minho
Escola de Engenharia

Ana Filipa Barros Duarte

Uma Proposta Semi-Automatizada para o Estabelecimento de Índices de Bem-Estar Cardíaco

Outubro de 2019

Universidade do Minho
Escola de Engenharia

Ana Filipa Barros Duarte

Uma Proposta Semi-Automatizada para o Estabelecimento de Índices de Bem-Estar Cardíaco

Dissertação de Mestrado
Mestrado em Engenharia de Sistemas

Trabalho efetuado sob a orientação do
Professor Doutor Orlando Manuel de Oliveira Belo

Outubro de 2019

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho

Atribuição-NãoComercial-SemDerivações
CC BY-NC-ND
https://creativecommons.org/licenses/by-nc-nd/4.0/

AGRADECIMENTOS

Um projeto de dissertação é o culminar de um ciclo de estudos de muita aprendizagem e de novos conhecimentos. No meu caso, para este projeto, pude contar com o apoio de várias pessoas que, de uma forma mais ou menos direta, contribuíram para o resultado final e para a produção do presente relatório. A todas elas devo-lhes a minha gratidão, por terem funcionado como o sistema de suporte à decisão deste projeto.

Desta forma, aproveito, em primeiro lugar, para agradecer ao meu orientador, o Professor Orlando Belo, por todos os ensinamentos que me transmitiu, não só no âmbito da dissertação, como também nas suas aulas, e por toda a atenção, disponibilidade, motivação e confiança que me deu para a concretização deste trabalho.

Não posso deixar de agradecer também à minha família, em particular aos meus pais, pelo apoio incondicional que sempre me deram ao longo do trabalho, do curso e da vida, e por incentivarem todas as minhas decisões.

Um obrigada especial, a todos os meus professores que me ensinaram tudo aquilo que hoje sei. Por último, aos meus colegas que comigo partilharam a vida escolar e aos meus amigos que compartilharam diferentes fases da minha vida deixo-lhes também o meu mais sincero agradecimento, por toda a ajuda e carinho recebidos. Não os posso referir nominalmente a todos, mas um pouco de cada um deles está também presente neste trabalho.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

Atualmente, as doenças cardiovasculares representam uma das maiores preocupações ao nível da Saúde, por se tratarem de uma das principais causas de morte nos países mais desenvolvidos. Além disso, muitas dessas mortes poderiam ser evitadas pela mudança de comportamentos no quotidiano ou pela toma de medicação preventiva em indivíduos propensos a desenvolverem este tipo de doenças. Deste modo, o principal esforço para a diminuição da sua incidência tem recaído na identificação dos fatores de risco e da sua influência e interação no resultado final, de se vir ou não a desenvolver esta doença. Neste contexto, a presente dissertação aborda a implementação de um sistema capaz de disponibilizar índices de bem-estar cardíaco, facilmente entendíveis, que funcionem como indutores para a adoção de hábitos de vida mais saudáveis, promovendo uma aposta contínua na prevenção deste tipo de doenças.

Para a concretização do sistema proposto, construiu-se, numa primeira etapa, um modelo preditivo, recorrendo-se a técnicas de Data Mining, que permitiu definir as contribuições e o relacionamento entre vários fatores de risco no surgimento de doenças cardiovasculares. Os resultados obtidos mostraram que o algoritmo *MultiLayer Perceptron* foi o que apresentou os melhores valores de acurácia e de sensibilidade (72.3% e 71.2%, respetivamente). Numa segunda fase, foram armazenados dados de utilizadores num sistema de *Data Warehousing* e calculados os seus índices, tendo-se por base o modelo preditivo selecionado. Além disso, de forma a disponibilizarem-se os resultados de um modo mais intuitivo para os *stakeholders*, foram criados, numa última etapa, *dashboards* direcionados para cada um deste tipo de utilizadores. Destes gráficos, retirou-se que o valor do índice global era de -1.73 e que, a nível individual, o seu valor pode ser ponderado em função dos cálculos passados. A nível visual também se demonstrou que os índices fornecem a informação individual, geral e geográfica de uma forma compacta e facilmente percetível.

Assim, concluiu-se que o sistema proposto pode funcionar como um sistema de apoio à decisão para os diversos intervenientes da área da Saúde e que um índice por dia e uma procura contínua por um estilo de vida mais saudável podem resultar numa forma eficiente de se manter as doenças cardíacas afastadas.

PALAVRAS-CHAVE

Cubo OLAP, Data Mining, Data Warehouse, Doença Cardiovascular, Índice de Bem-Estar

ABSTRACT

Nowadays, heart diseases represent one of the major health concerns, as they are one of the leading causes of death in developed countries. In addition, many of these deaths could be prevented by changing daily behaviour or taking preventive medication in individuals more likely to develop this type of diseases. In this way, the main efforts to reduce their incidence have been made in order to identify risk factors and their influence and interaction in the final outcome. In this context, the present dissertation addresses the implementation of a system capable of providing easily understandable cardiac well-being indexes that act as inducers for the adoption of healthier lifestyle habits, promoting a continuous progress in the prevention of this type of diseases.

For the implementation of the proposed system, a predictive model was built, in a first step, using Data Mining techniques, which allowed defining the contributions and the relationship between various risk factors in the onset of cardiovascular diseases. The results showed that MultiLayer Perceptron was the best algorithm in terms of accuracy and sensitivity values (72.3% and 71.2%, respectively). In a second phase, users' data were stored in a Data Warehousing system and their indexes were calculated based on the selected predictive model. Besides that, in order to make the results more easily available to stakeholders, dashboards were created in a last step for each one of this type of users. From these graphs, global index value was found to be -1.73 and that, on an individual basis, this value can be weighted considering past calculations. At the visual level it has also been shown that indexes provide individual, general and geographic information in a compact and easily perceptible manner.

Thus, it was concluded that the suggested system can act as a decision support system for the several Health actors and that an index a day can keep heart diseases away.

KEYWORDS

Data Mining, Data Warehouse, Heart Disease, OLAP Cube, Well-being Index

ÍNDICE

1. Introdução 1
1.1 Contextualização 1
1.2 Motivação 3
1.3 Objetivos 4
1.4 Organização da Dissertação 5
2. As Diferentes Panorâmicas das Doenças Cardiovasculares 7
2.1 O Estudo de Framingham 7
2.2 Os Atuais Simuladores de Risco Cardíaco 9
2.3 As Técnicas de DM na Previsão das Doenças Cardiovasculares 11
2.4 A Importância da Promoção de uma Monitorização Contínua 14
3. A Desmistificação dos Padrões Ocultos 17
3.1 Compreensão do Negócio 17
3.2 Compreensão dos Dados 19
3.3 Preparação dos Dados 24
3.4 Modelação 28
3.5 Avaliação 50
3.6 Implementação 55
4. O Processo de Armazenamento dos Dados 58
4.1 Planeamento e Gestão do SDW 58
4.2 Levantamentos dos Requisitos 59
4.3 Modelação Dimensional 61
4.4 Fontes de Informação 63
4.5 Implementação do SDW 70
4.5.1 Implementação dos Esquemas Físicos 71
4.5.2 Aspetos Gerais da Implementação 72
4.5.3 ETL das Tabelas de Dimensão 76
4.5.4 ETL da Tabela de Factos 85
4.5.5 Validação e Considerações Finais do SDW Implementado 90
5. Uma Visualização Interativa dos Resultados 95
5.1 A Construção do Cubo OLAP 95

5.2 *Dashboards* para Utilizadores Individuais e Profissionais de Saúde 97
5.3 *Dashboards* para Fins Estatísticos 98
5.4 *Dashboard* do Histórico dos Utilizadores 103
6. Conclusões e Trabalho Futuro 104
Referências Bibliográficas 107
Anexo I – Arquitetura das Redes Neuronais MLP (Exemplo Cenário I) 111
Anexo II – Simulações Efetuadas para a Otimização dos Parâmetros das Técnicas de DM (Cenário I) 112
Anexo III – Simulações Efetuadas para a Otimização dos Parâmetros das Técnicas de DM (Cenário II) 123
Anexo IV – Modelos Gerados pelas Técnicas RF, NB e MLP (Cenário I) 134
Anexo V – Caracterização das Tabelas de Dimensão do DW 135
Anexo VI – Caracterização da Tabela de Factos do DW 138
Anexo VII – Perfis de Dados 139
Anexo VIII – *Triggers* e *Stored Procedures* 141
Anexo IX – Extrato de *Emails* de Confirmação de Sucesso de DM e ETL 143

ÍNDICE DE FIGURAS

Figura 1 - Esquematização do modelo proposto para o sistema dos índices de bem-estar. 5
Figura 2 - Metodologia seguida por Kim e Kang (2017). 14
Figura 3 - Composição e distribuição inicial do *dataset* de DM. 20
Figura 4 - Estatísticas básicas relativas a cada um dos atributos do *dataset*. 22
Figura 5 - Extrato da visualização gráfica entre pares de atributos. 23
Figura 6 - Relacionamento da pressão arterial alta (eixo horizontal) com a pressão arterial baixa (eixo vertical). 23
Figura 7 - Arquitetura do tratamento de dados para o processo de DM (Etapa 1). 24
Figura 8 - Condições, em Java, para assegurar a coerência dos dados referentes ao tabagismo. 25
Figura 9 - Código Java para contagem de campos omissos por registo. 26
Figura 10 - Arquitetura do tratamento de dados para o processo de DM (Etapa 2). 26
Figura 11 - Planeamento da metodologia a seguir para testar a qualidade e a validade do modelo. 33
Figura 12 - Arquitetura da modelação do processo de DM. 34
Figura 13 - Excerto da árvore gerada pelo algoritmo J48 (cenário II). 45
Figura 14 - Implementação, no Spoon, para a seleção do melhor modelo para o caso em estudo. 48
Figura 15 - *Scores* relativos a cada um dos modelos analisados. 49
Figura 16 - Curva de correspondência entre o grau de risco de DCV e o valor do índice. 51
Figura 17 - Previsões da cor do índice associado a cada registo, de acordo com a classe. 52
Figura 18 - Perfis-tipo para comparação dos índices calculados com os dos simuladores *online*. 53
Figura 19 - Possível implementação do modelo desenvolvido. 56
Figura 20 - Diagrama de Gantt do planeamento do projeto de SDW. 59
Figura 21 - Esquema conceptual do caso em estudo. 63
Figura 22 - Criação das tarefas de *profiling* dos dados das Fontes 1 e 2. 67
Figura 23 - Perfis de estatísticas por coluna, referentes à Fonte 1. 68
Figura 24 - Perfis de estatísticas por coluna, referentes à Fonte 2. 69
Figura 25 - Mecanismo de migração dos dados. 70
Figura 26 - Estruturas implementadas na DSA. 71

Figura 27 - Estruturas implementadas no DW. 71
Figura 28 - Programação do início de execução da *job* relativa ao DM. 72
Figura 29 - *Job* principal responsável pelo desencadeamento do processo de DM. 73
Figura 30 - *Job* principal responsável pelo desencadeamento do processo de ETL. 73
Figura 31 - *Job* relativa ao processo de DM. 74
Figura 32 - Esquema BPMN da metodologia de povoamento do DW. 74
Figura 33 - Primeira *job* incorporada na *job ETL*. 75
Figura 34 - Segunda *job* incorporada na *job ETL*. 75
Figura 35 - Constituição da *job DimUtilizador* 76
Figura 36 - Esquema BPMN da metodologia de extração dos dados para o povoamento da *DimUtilizador* 78
Figura 37 - Esquema BPMN da metodologia de transformação dos dados para o povoamento da *DimUtilizador*. 78
Figura 38 - Implementação, no Spoon, da transformação relativa ao processo de limpeza da *DimUtilizador* 79
Figura 39 - Constituição da *job Carregar Utilizador*. 80
Figura 40 - Esquema BPMN do processo de criação das SK para o povoamento da *DimUtilizador* 80
Figura 41 - Implementação, no Spoon, da transformação relativa ao processo de carregamento da *DimUtilizador*. 83
Figura 42 - Processo de *Surrogate Key Generator* relativo à *DimDistrito*. 84
Figura 43 - Metodologia para o povoamento da *DimCalendário*. 84
Figura 44 - Constituição da *job preTF*. 85
Figura 45 - Implementação, no Spoon, da transformação relativa ao processo de extração dos dados da Fonte 2 da *TFBemEstar*. 86
Figura 46 - Condições, em Java, para garantir a validade máxima de um ano entre as datas do índice e das análises e que a data do índice é superior à das análises. 89
Figura 47 - Constituição da *job TFBemEstar*. 89
Figura 48 - Processo de *Surrogate Key Pipeline*. 90
Figura 49 - Excerto da dimensão *Utilizador* após o primeiro povoamento. 91
Figura 50 - Excerto da tabela de *log* relativa às tarefas de extração da Fonte 2. 92
Figura 51 - Excerto da tabela de quarentena relativa à tabela de factos. 92
Figura 52 - Conteúdo da tabela de auditoria após um conjunto de modificações à Fonte 1... 93
Figura 53 - Excerto da dimensão *Utilizador* após o povoamento incremental. 93

Figura 54 - Histórico do utilizador após o povoamento incremental. 93
Figura 55 - Instrução, em linguagem MDX, para a determinação do membro calculado *LinRegPoint*. 96
Figura 56 - Instrução, em linguagem MDX, para a determinação do membro calculado *ÍndicePonderado*. 96
Figura 57 - Instrução, em linguagem MDX, para a determinação do membro calculado *ÍndiceGlobal*. 96
Figura 58 - *Dashboards* e indicadores relativos ao registo mais recente do utilizador de *id* 2961. 97
Figura 59 - *Dashboards* e indicadores relativos ao registo do dia 14/12/2018 do utilizador de *id* 2961. 98
Figura 60 - *Dashboards* para avaliação dos valores de índice globais. 99
Figura 61 - *Dashboards* para avaliação dos valores de índice globais em Viana do Castelo. 101
Figura 62 - Mapa de Portugal colorido em função dos valores dos índices de cada distrito. 102
Figura 63 - Tabela de histórico do utilizador com *id* 4. 103

ÍNDICE DE TABELAS

Tabela 1 - Exemplos de simuladores online para o cálculo do risco cardíaco 10
Tabela 2 - Distribuição dos dados dos atributos com valores extremos de acordo com intervalos 21
Tabela 3 - Exemplos de estudos desenvolvidos para previsão de DCV 29
Tabela 4 - Valores a otimizar nos parâmetros do algoritmo J48 36
Tabela 5 - Valores a otimizar nos parâmetros do algoritmo RF 36
Tabela 6 - Valores a otimizar nos parâmetros do algoritmo NB 37
Tabela 7 - Valores a otimizar nos parâmetros do algoritmo KNN 38
Tabela 8 - Valores a otimizar nos parâmetros do algoritmo MLP 39
Tabela 9 - Sumário da configuração dos parâmetros para as técnicas em estudo 44
Tabela 10 - Matriz de confusão relativa a cada uma das técnicas em estudo para a previsão de DCV 46
Tabela 11 - Quadro comparativo entre as fases de treino e validação e a de teste dos modelos 49
Tabela 12 - Percentagem de risco de desenvolvimento de DCV a 10 anos pelos simuladores *online* 54
Tabela 13 - Sumário das fontes candidatas 64
Tabela 14 - Conteúdo das fontes candidatas 65
Tabela 15 - Mapeamento entre as fontes de dados e o DW 70
Tabela 16 - Validação dos dados 78
Tabela 17 - Extrato da tabela de *log* que é criada na DSA 86

LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

BPMN	*Business Process Model and Notation*
CDC	*Change Data Capture*
CRISP-DM	*CRoss Industry Standard Process for Data Mining*
DCV	Doença Cardiovascular
DM	*Data Mining*
DSA	Área de Retenção
DT	*Decision Trees*
DW	*Data Warehouse*
ETL	Extração, Transformação e Carregamento
FHS	*Framingham Heart Study*
FRS	*Framingham Risk Score*
KNN	*K-Nearest Neighbors*
MLP	*MultiLayer Perceptron*
NB	*Naïve Bayes*
NN	*Neural Network*
OLAP	*Online Analytical Processing*
OMS	Organização Mundial de Saúde
RF	*Random Forest*
SAD	Sistema de Apoio à Decisão
SCD	*Slowly Changing Dimension*
SDW	Sistema de *Data Warehousing*
SI	Sistema de Informação
SVM	*Support Vector Machine*
SK	*Surrogate Key*

"A grande originalidade não é dizer coisas novas, mas ser novo diante das coisas velhas." (Vergílio Ferreira, in Conta-Corrente 3)

1. INTRODUÇÃO

1.1 Contextualização

Garantir cuidados de saúde com qualidade é imprescindível para que se possa viver mais e melhor. É com base neste mote que, atualmente, a Saúde representa uma das áreas de maior investimento a nível científico, social e mesmo pessoal. Uma parcela significativa dos crescentes níveis de esperança média de vida que se registam nos países desenvolvidos deve-se à procura contínua pelos melhores cuidados de saúde. A qualidade na Saúde deriva, essencialmente, de três vertentes: do grau de desenvolvimento da ciência médica, da disponibilidade do conhecimento relevante e da acessibilidade aos melhores meios de diagnóstico e de tratamento.

No campo do desenvolvimento da Medicina, assistiu-se, sobretudo após a segunda guerra mundial, a uma sistematização da procura de conhecimento, com a realização de experiências que consistiam no acompanhamento regular dos doentes, durante longos períodos de tempo. Nessa época, a investigação médica tinha como objeto de estudo as relações de causa-efeito apenas com recurso à monitorização dos pacientes e ao tratamento estatístico dos dados obtidos, sendo essas relações apuradas através da análise de variáveis isoladas. Com o decorrer do tempo, a Medicina começou a incorporar a influência de diferentes variáveis e a examinar o tipo de interação existente entre elas. À medida que mais parâmetros iam sendo adicionados, o processo de análise tornava-se gradativamente mais complexo, conduzindo a investigações cada vez mais limitadas e difíceis de compreender e de interpretar. Neste contexto, mais recentemente, a Medicina passou a integrar também na investigação ferramentas informáticas e a tirar partido das suas vantagens ao nível do armazenamento, do tratamento e da análise dos dados. Nos últimos anos, o desenvolvimento da Informática tem tido um crescimento exponencial, desenvolvendo-se *software* próprio para o armazenamento e para o fornecimento de dados acessíveis, fidedignos e preparados para as necessidades de cada utilizador. Além disso, no que diz respeito ao tratamento e à análise dos dados, a Informática também propiciou o desenvolvimento de novas abordagens, com particular relevo para a utilização dos algoritmos de *Data Mining* (DM). Estes algoritmos, que aliam diferentes técnicas estatísticas e heurísticas, têm vindo a revolucionar o paradigma existente, ao possibilitarem análises complexas aos

dados, de uma forma eficiente e automatizada, com o propósito de se retirar conhecimento a partir deles. Desta forma, os dados que se geram no quotidiano têm vindo a assumir cada vez mais valor e quanto mais dados forem recolhidos, maior será, consequentemente, o conhecimento descoberto. Assim, estas técnicas de DM mostram que, muitas vezes, uma visão do passado pode ser fundamental para se perceber o futuro. Citando Vergílio Ferreira, "a grande originalidade não é dizer coisas novas, mas ser novo diante das coisas velhas". Neste caso, a grande originalidade da ciência não passa por olhar para o futuro, mas sim por, diante dos dados que já existem, tentar encontrar padrões e correspondências que permitam apurar as causas e, a partir delas, extrair conhecimento.

Quanto à segunda vertente, relativa ao conhecimento dos avanços médicos, importa que as novas descobertas, sobretudo o conhecimento que se relaciona com o estilo de vida de cada um e o que isso representa para a sua saúde, sejam não só apresentadas aos profissionais de saúde, como também difundidas por toda a população. A este nível, uma vez mais, a Informática assume um importante papel, ao possibilitar a divulgação do conhecimento médico, de uma forma simplista, à generalidade da população. Deste modo, uma percentagem significativa da sociedade age de acordo com este conhecimento, de uma forma mais conscienciosa e sensibilizada, ajustando os seus hábitos e comportamentos, no sentido de preservar a sua saúde. Outro dos aspetos inerentes a esta vertente prende-se com a disponibilização desta informação para apoio à decisão. Organismos públicos, decisores de políticas públicas e outros *stakeholders* da área da Saúde apreendem e incutem nos seus processos deliberativos a informação que lhes é apresentada.

Por sua vez, a vertente relativa aos meios de diagnóstico e de tratamento está diretamente relacionada com o nível de vida das comunidades e com o custo que podem suportar. Países mais evoluídos e com maiores rendimentos apresentam, por norma, mais e melhores meios e, consequentemente, melhores cuidados de saúde e maiores valores de esperança média de vida. Neste campo, a Informática aplicada à Medicina pode apresentar vantagens consideráveis ao permitir, por exemplo, compatibilizar as técnicas de DM com a conceção de Sistemas de Apoio à Decisão (SAD) que funcionem como alternativas fiáveis aos equipamentos complementares de diagnóstico existentes. Assim, estes SAD, além de permitirem melhorar os cuidados de saúde prestados e reduzir o número necessário de exames de diagnóstico, possibilitam também a redução dos custos associados à compra e à manutenção de equipamentos, sobretudo nos países com menos recursos económicos.

1.2 Motivação

Como se depreende, a utilização de *software* informático apropriado para aplicação na área médica pode resultar em oportunidades de se criar mais por menos: mais qualidade por menos custos.

Nas populações atuais mais desenvolvidas, as principais preocupações com os cuidados de saúde centram-se nas doenças que causam a maioria das mortes registadas e que poderiam ser prevenidas. Nesse contexto, as doenças cardiovasculares (DCV) representam um dos principais focos de atenção, uma vez que apresentam uma das taxas de mortalidade mais elevadas e têm uma grande margem de melhoria ao nível da sua prevenção. Do ponto de vista clínico, estas doenças têm sido alvo de grande investigação e alguns dos seus principais fatores de risco podem ser modificáveis por via dos hábitos de vida ou por via medicamentosa. Assim, neste caso, as principais medidas para a redução do risco de DCV passam pelo incentivo à adoção de estilos de vida mais saudáveis e pela promoção do acompanhamento médico.

O caso das DCV é um dos exemplos em que a utilização de diversas ferramentas informáticas pode servir como um importante auxílio para a melhoria dos cuidados de saúde. A partir destas ferramentas podem ser, por exemplo, identificados os principais fatores de risco e a sua correlação no desenvolvimento de DCV. Desta forma, possibilita-se uma análise diferenciada e adaptada ao tipo de perfil de cada paciente, que permite estimar, de acordo com valores probabilísticos, o risco de progressão de DCV associado a cada um deles. Consoante a triagem efetuada, os meios de diagnóstico a serem utilizados podem ser adaptados a cada um dos perfis, obtendo-se, assim, uma maior personalização e, em consequência disso, um tratamento mais eficaz com menores custos. Deste modo, abrir-se o leque de opções e inovar-se nas estratégias de atuação permite que, em situações de cariz clínico, possam existir menos incertezas no diagnóstico médico.

Outro dos aspetos em que a Informática se pode destacar, ainda no âmbito das DCV, é na forma de comunicação e de transmissão do conhecimento útil acerca deste tipo de doenças. Um exemplo disso é a possibilidade de se gerarem e disponibilizarem gráficos intuitivos à população, que permitam acompanhar em tempo real cada um dos indivíduos, e alertá-los nos casos de risco elevado de doença. Outro exemplo prende-se com o facto de esta facilidade em transmitir informação poder fomentar a adoção de políticas e campanhas de saúde adaptadas e corretamente direcionadas a cada região e modo de vida, incidindo nas debilidades intrínsecas de cada área do país.

No fundo, neste contexto, o motivo principal para se recorrer a técnicas informáticas é o facto de se proporcionar mais tempo de qualidade para todos e menos tempo para as DCV.

1.3 Objetivos

O objetivo principal da presente dissertação passa pela disponibilização de um sistema capaz de medir os índices de bem-estar cardíaco dos seus utilizadores. Neste caso, estes índices deverão ter por base uma escala de cores, que possibilite uma avaliação mais intuitiva e inequívoca. Deste modo, pretende-se que os índices criados assumam valores sólidos e que sejam capazes de detetar todos os casos de elevado risco de desenvolvimento de DCV. Assim, o sistema proposto tem o intuito de promover a adoção, por parte dos utilizadores, de comportamentos que não prejudiquem a sua saúde cardiovascular, mantendo-os focados no sentido de melhorarem, cada vez mais, o valor do seu índice. Para isso, o sistema deverá permitir, para cada um dos utilizadores, registar e guardar os valores diários dos seus índices e, para efeitos de consulta, deverão refletir um valor ponderado com os anteriores para que, em casos de medições atípicas, o valor do índice não seja inflacionado/desinflacionado. Com isso, pretende-se criar um SAD vocacionado para utilizadores comuns e para profissionais de saúde, que visa prevenir e reduzir significativamente a incidência de DCV no país, e ainda possibilite a identificação das regiões críticas, em que, em média, os indivíduos apresentam piores valores globais.

Com esta finalidade em vista, definiu-se um conjunto de objetivos parciais que concorrem para a sua concretização:

- desenvolver modelos preditivos que sejam criados e avaliados de uma forma automática, e que garantam, em simultâneo, uma elevada taxa de acerto e uma baixa identificação errónea de pessoas doentes;
- elaborar índices de bem-estar cardíaco a partir do melhor modelo preditivo encontrado;
- criar um sistema de armazenamento DW/OLAP que congregue toda a informação relativa aos dados pessoais, clínicos e índices de bem-estar de utilizadores comuns;
- disponibilizar, a partir do sistema criado, os valores dos índices de uma forma ponderada com os valores históricos dos utilizadores, podendo ser agregados de acordo com diferentes perspetivas, que facilitem a sua compreensão;

- distribuir os valores médios dos índices em cada região através de mapas coloridos, que associem a cada distrito um valor e uma escala de risco, que varie desde o risco baixo (cor verde) até ao risco elevado (cor vermelha).

De um ponto de vista esquemático, com este trabalho pretendeu-se implementar o sistema indicado na **Figura 1**.

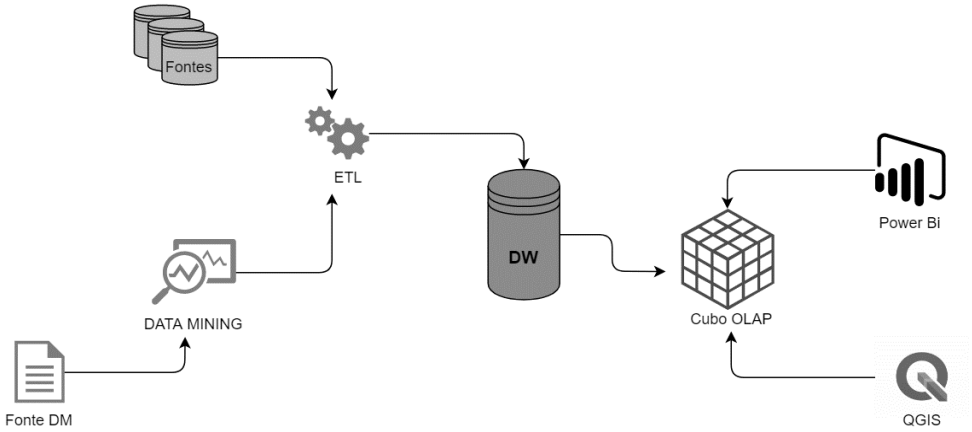


Figura 1 - Esquematização do modelo proposto para o sistema dos índices de bem-estar.

Desta forma, ao cumprirem-se os objetivos expostos espera-se contar com a aderência por parte dos utilizadores, por forma a contribuir-se com soluções que mitiguem parte da falta de informação existente acerca das DCV e apelem à sua prevenção.

1.4 Organização da Dissertação

De um ponto de vista estrutural, o presente trabalho de dissertação está dividido e integrado em seis capítulos principais. A descrição sucinta do conteúdo de cada um destes capítulos é a que se descreve abaixo, à exceção da presente secção que diz respeito à *Introdução*, que não foi incluída:

- **Capítulo 2 – As Diferentes Panorâmicas das Doenças Cardiovasculares.**

  Este capítulo inicia-se de uma forma mais genérica com o estado do conhecimento atual relativo às DCV, evidenciando os principais fatores de risco conhecidos e identificando, cronologicamente, o período da sua descoberta. Apresentados estes fatores, são também abordados alguns dos simuladores existentes criados com o propósito de informar acerca do risco cardíaco. Ainda neste capítulo, descrevem-se diferentes estudos realizados com recurso a técnicas de DM, efetuados com o intuito de, a partir das principais causas encontradas, conseguir prever-se o surgimento deste tipo de doenças e os seus principais

fatores de risco. No final, destaca-se a importância de uma monitorização contínua e indicam-se alguns estudos que se relacionam com a implementação de *dashboards*.

- **Capítulo 3 – A Desmistificação dos Padrões Ocultos.**

  Nesta secção do trabalho apresenta-se uma proposta de implementação de um sistema automatizado para o cálculo de índices de bem-estar cardíaco, que se relaciona com a criação de modelos preditivos baseados em técnicas de DM. Assim, todo o processo de tratamento de dados e de escolha de atributos, apesar de não dispensar uma validação periódica realizada por analistas, apresenta a particularidade de poder ser realizado de forma automática. Além disso, a tarefa da seleção do modelo mais adequado, que corresponde àquele que maximiza a acurácia e minimiza as identificações erradas de pessoas que sofram de DCV, também ela pode ser efetuada automaticamente.

- **Capítulo 4 – O Processo de Armazenamento dos Dados.**

  No capítulo 4 é descrita a metodologia considerada para a criação e para o povoamento do Sistema de *Data Warehousing* (SDW) proposto, detalhando-se as várias etapas e o modo de as concretizar. No final do capítulo, são ainda descritos os testes efetuados no sentido de se aferir o correto funcionamento do SDW.

- **Capítulo 5 – Uma Visualização Interativa dos Resultados.**

  Esta secção relaciona-se com a construção de um cubo OLAP e com a apresentação e discussão dos principais resultados obtidos pela implementação do sistema DW/OLAP. Assim, nesta etapa, mostram-se os gráficos construídos para os índices, tendo em conta as perspetivas estudadas, de modo a realçar-se a sua importância em função de diferentes aspetos de análise, como o valor do índice geral ou do índice por utilizador ou por distrito.

- **Capítulo 6 – Conclusões e Trabalho Futuro.**

  Na última etapa do trabalho, foram sintetizadas as principais conclusões, tendo em atenção os objetivos iniciais propostos. Adicionalmente, foram enumeradas as principais vantagens e benefícios que um SAD deste tipo pode proporcionar aos seus utilizadores e *stakeholders* da área da Saúde. Além disso, foram descritas as limitações encontradas e sugeridas formas de as ultrapassar em trabalhos futuros, assim como foram também referidas áreas que, não se integrando diretamente neste trabalho, poderiam resultar num potencial complemento para aplicações reais.

2. AS DIFERENTES PANORÂMICAS DAS DOENÇAS CARDIOVASCULARES

Na década de 40 não existiam medidas preventivas nem quaisquer tratamentos para os pacientes com DCV. Na época, estas doenças eram encaradas como sinónimo de uma morte precoce, sem que nada se pudesse fazer para se mudar este destino. No final desse período, começaram a ser descobertos os primeiros fatores de risco e, no ano de 1948, iniciou-se um estudo que ainda hoje vigora, no sentido de se confirmarem e descobrirem quais os fatores de risco que mais conduzem ao desenvolvimento de DCV. Deste modo, com base nos parâmetros encontrados, começaram a surgir sistemas que retornavam valores de risco associados à probabilidade de indivíduos em específico virem a ter DCV.

2.1 O Estudo de Framingham

O estudo epidemiológico pioneiro ao nível das DCV foi iniciado em 1948, em Framingham – Massachusetts, EUA – por se tratar de uma cidade muito ligada à investigação médica. Neste estudo, participaram 5209 indivíduos residentes na cidade, entre os 28 e os 62 anos, com uma média de 44 anos, dos quais cerca de pouco mais de metade eram do sexo feminino (Mahmood *et al.*, 2014; Andersson *et al.*, 2019).

Mais tarde, Kannel *et al.* (1961) identificaram o sexo, a idade, a pressão arterial alta, a hipertensão, o colesterol total e a hipertrofia ventricular esquerda como sendo importantes parâmetros relacionados com a manifestação de DCV. Deste modo, no ano de 1961, o *Framingham Heart Study* (FHS) indicou que os indivíduos mais propensos a padecer deste tipo de doenças eram os do sexo masculino, idosos, com valores elevados de pressão arterial alta e de colesterol, e ainda com anormalidades específicas detetadas nos registos dos eletrocardiogramas.

Estudos posteriores, levados a cabo entre 1962 e 1964, comprovaram ainda a existência de uma relação direta entre o facto de se ser fumador e o aumento do risco de progressão de DCV. Já no ano de 1967, verificou-se que a atividade física era inversamente proporcional ao grau de risco de desenvolvimento destas doenças e, em anos posteriores, foram acrescentados os fatores de risco associados à obesidade, à fibrilação auricular e à diabetes *mellitus* (Andersson *et al.*, 2019).

Posteriormente, Wilson *et al.* (1998) publicaram uma metodologia para a determinação do risco, a dez anos, de se vir a sofrer algum tipo de evento cardíaco grave, que serviu de base a todo o conhecimento atual existente. A metodologia desenvolvida baseava-se na atribuição de pontuações associadas aos atributos idade, colesterol total (em alternativa, também se

poderia considerar o colesterol LDL), colesterol HDL, pressão arterial sistólica (alta) e diastólica (baixa), diabetes e hábitos tabágicos. De acordo com este método, estes pontos eram somados e o *Framingham Risk Score* (FRS) correspondia a uma percentagem de risco tabelada, que estava associada aos pontos totais obtidos. Assim, este mecanismo de cálculo possibilitou que os médicos, através de uma escala de cores, tivessem à sua disposição um sistema apto para prever e classificar a probabilidade de contração de DCV em *Risco Muito Baixo*, *Risco Baixo*, *Risco Moderado*, *Risco Alto* e *Risco Muito Alto*.

Apesar dos já consideráveis contributos do FHS, o estudo iniciou uma nova etapa em 2002, com o recrutamento de novos participantes, filhos de participantes anteriores, de modo a analisar-se também a influência dos fatores genéticos. Além disso, o número de intervenientes no estudo foi também equilibrado de acordo com a etnia, uma vez que a população original de Framingham era sobretudo branca de descendência europeia. (Mahmood *et al.*, 2014)

Decorridos mais de 70 anos após o seu início, o FHS continua ativo e é hoje reconhecido como o mais importante estudo mundial sobre DCV, sendo o que mais contribuiu para a determinação dos seus principais fatores de risco. Dado o seu sucesso, vários países adaptaram o FRS à sua realidade, utilizando a mesma metodologia de cálculo. Um dos exemplos destes países é Espanha, em que Marrugat *et al.* (2003) ajustaram o FRS à população do país.

Tendo constatado que as equações do FRS original sobrestimavam o risco de se sofrer de DCV em países onde existe uma baixa taxa de incidência destas doenças, os autores adaptaram o cálculo da probabilidade de surgimento de eventos cardíacos graves a dez anos para a população espanhola. Este estudo teve, assim, como principal objetivo a calibração das tabelas de cálculo de risco cardíaco à realidade espanhola, com base no FHS, sem ter sido necessária a criação de modelos preditivos próprios. A título de exemplo, nesta investigação baseada em pacientes da região espanhola de Girona, observou-se que o risco associado ao colesterol HDL variava significativamente entre esta população e a de Framingham. Para a população espanhola, níveis de HDL inferiores a 35 mg/dL aumentavam o risco em cerca de 50% e, pelo lado contrário, níveis superiores a 60 mg/dL diminuíam-no em, aproximadamente, 50%. Com base nos resultados apurados, foi construído um conjunto de tabelas idênticas às do FRS, que fazem corresponder cada fator de risco a uma determinada pontuação. Do conjunto das pontuações de todas as tabelas, obtém-se o grau de risco adaptado para a população espanhola.

De forma análoga, as tabelas do FRS foram ajustadas para outras populações-alvo, seguindo sempre a mesma metodologia.

2.2 Os Atuais Simuladores de Risco Cardíaco

Uma consequência direta da determinação do FRS e de outros estudos desenvolvidos no âmbito dos fatores de risco de DCV é a possibilidade de se implementarem simuladores que permitem calcular o risco cardíaco. Atualmente, existem diversas ferramentas disponibilizadas para este fim, que possibilitam esta avaliação.

Um dos simuladores capazes de determinar o grau de risco de um evento cardiovascular grave é o que é providenciado pela *American Heart Association*[1]. Esta associação dispõe de uma ferramenta *online* e permite estimar este risco a partir do preenchimento obrigatório dos campos relativos à idade, colesterol total, colesterol LDL, colesterol HDL, pressão arterial alta, pressão arterial baixa, sexo e etnia. Posteriormente, deve também especificar-se se o indivíduo é, ou não, um atual fumador. Além destes, devem também ser assinaladas, quando aplicáveis, as opções referentes ao historial clínico no que diz respeito à toma de estatinas, aspirinas ou de medicação para a hipertensão, existência de diabetes, ataque cardíaco, AVC grave ou ligeiro, angina, doença arterial periférica e registo de outras DCV. Assim, depois de detalhados todos os campos, é calculado e apresentado o risco correspondente de se vir a ter uma complicação cardiovascular grave, como um ataque cardíaco ou um AVC, a dez anos.

Outro instrumento que pode ser utilizado para a determinação do risco cardíaco é o ASSIGN[2], que foi desenvolvido na Universidade de Dundee, na Escócia, em 2006. Este simulador é direcionado para indivíduos residentes na Escócia e acrescenta os atributos dos antecedentes familiares e do índice escocês de privação múltipla (que pode ser inferido pelo código postal de residência dos utilizadores). Este índice é aplicado apenas a cidadãos escoceses e refere-se à qualidade e ao desenvolvimento global da área de residência dos cidadãos em termos de educação, saúde, crime e empregabilidade, por exemplo, e assume que indivíduos de uma mesma área (código postal) têm um *status* social semelhante entre si. Os restantes fatores de risco considerados são os que se relacionam com a idade, sexo, diabetes, número de cigarros diário fumados, pressão arterial alta, colesterol total e colesterol HDL.

Em Itália, o projeto Cuore[3], fundado em 1998 pelo Instituto Nacional da Saúde italiano, dispõe de um simulador *online* que calcula o risco de se sofrer de um evento cardiovascular grave com base no sexo, idade, hábitos tabágicos, pressão arterial sistólica, colesterol total,

[1] Simulador *American Heart Association* disponível em: https://ccccalculator.ccctracker.com/
[2] Simulador ASSIGN disponível em: http://www.assign-score.com/estimate-the-risk/
[3] Simulador Projeto Cuore disponível em: http://www.cuore.iss.it/sopra/calc-rischio_en.asp

colesterol HDL, existência de diabetes e de hipertensão, e está adaptado para a população italiana.

Todos estes simuladores retornam um valor percentual relativo ao risco de eventuais complicações cardíacas graves e o valor obtido pode incluir-se numa das seguintes categorias:

- **Risco Baixo:** valores de risco inferiores a 5%.
- **Risco Leve:** valores de risco entre 5 e 7.5%.
- **Risco Intermédio:** valores de risco entre 7.5 e 20%.
- **Risco Elevado:** valores de risco superiores a 20%.

Além destes, existem ainda outros simuladores também disponíveis *online* e a **Tabela 1** reúne alguns dos mais conhecidos.

Tabela 1 - Exemplos de simuladores online para o cálculo do risco cardíaco

Designação do Simulador	Hiperligação
QRISK®3-2018	https://qrisk.org/three
ASCVD	http://tools.acc.org/ASCVD-Risk-Estimator-Plus
FHS Cardiovascular Disease (risco a 10 anos)	https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk
FHS Cardiovascular Disease (risco a 30 anos)	https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-30-year-risk
Reynolds Risk Score	http://www.reynoldsriskscore.org
Globorisk	http://www.globorisk.org/calc/officeform
Maggic	http://www.heartfailurerisk.org/
U-Prevent	https://www.u-prevent.com/en-GB/Calculators/Menu

Uma das características que os simuladores atuais apresentam prende-se com o facto de utilizarem algoritmos “rígidos”, cuja metodologia de cálculo permanece a mesma a longo prazo e despreza o relacionamento entre os fatores de risco ou, pelo menos, não é capaz de ter em conta relacionamentos entre eles que requeiram uma maior complexidade. Neste sentido, as técnicas de DM podem possibilitar ultrapassar estas limitações, pelo facto de permitirem gerar frequentemente novos algoritmos de cálculo de uma forma eficiente, além de admitirem a criação de métodos de cálculo mais complexos.

2.3 As Técnicas de DM na Previsão das Doenças Cardiovasculares

Os algoritmos de DM podem surgir, assim, como alternativas eficazes e complementares aos métodos tradicionais de cálculo de risco de complicações cardiovasculares críticas, capazes de fomentarem melhores serviços de saúde e menores custos envolvidos. Desta forma, as suas grandes vantagens relacionam-se com a criação de modelos mais complexos e sofisticados, em que a interação entre fatores de risco é tida em conta, e com a possibilidade de serem rapidamente executados, mesmo perante grandes conjuntos de dados. Além disso, em contraste com os métodos convencionais de cálculo de risco cardíaco, estas técnicas dispensam processos de recolha de dados tão dispendiosos.

Do ponto de vista da utilização das técnicas de DM, aplicadas ao nível das DCV, Palaniappan e Awang (2008) desenvolveram um sistema de previsão de doenças baseado nas técnicas *Decision Trees* (DT), *Naïve Bayes* (NB) e *Neural Networks* (NN). Os autores pretenderam proporcionar um sistema eficaz que permitisse, por um lado, uma identificação correta das situações associadas a um maior risco de ocorrência de DCV, para melhorar os serviços prestados e, por outro lado, que diminuísse os custos inerentes ao diagnóstico de DCV. O ficheiro de dados que utilizaram como suporte ao estudo foi o *dataset Cleveland Heart Disease*, que era constituído por 909 registos médicos que, por sua vez, continham informação referente a 15 atributos clínicos.

De uma forma global, verificou-se que o algoritmo NB foi o que melhor identificou corretamente as pessoas doentes (180) e o que menos doentes identificou incorretamente (28). Em relação aos indivíduos saudáveis, a técnica DT foi a que melhor os identificou (219) e foi também a que menos pessoas saudáveis previu como doentes (27).

Uma outra das etapas do projeto consistiu na identificação de padrões e relacionamentos entre fatores de risco, tendo-se como objetivo a determinação da probabilidade associada à ocorrência de uma DCV, de acordo com determinados perfis pessoais e clínicos pré-definidos. Neste caso, os resultados mostraram que a técnica NB foi a que apresentou uma maior capacidade preditiva, seguida da NN e, em último lugar, ficou a DT. No entanto, os valores preditivos foram todos próximos entre si e observou-se que a diferença da percentagem de acerto entre as técnicas NB e DT não era superior a 2%.

Através do seu estudo, os autores também constataram que, consoante a técnica utilizada, os atributos apresentavam diferentes relevâncias para a determinação das previsões. Por exemplo, a técnica NB considerou o fator "*chest pain type* = 4" como o mais relevante para a previsão, enquanto o algoritmo NN apontou para o "*oldpeak*", com os valores entre 3.05 e 3.81,

como o mais importante. Além disso, também salientaram o facto de os resultados obtidos através da DT serem os mais facilmente entendíveis, referindo que, no caso desta técnica, a regra mais relevante para a determinação do risco de DCV era "*chest pain type* = 4 e CA = 0 e *exang* = 0 e 146.362 <= *trest blood pressure* < 158.036".

Karaolis *et al.* (2010) conduziram outro estudo baseado na utilização de DT para a avaliação dos fatores de risco associados aos eventos enfarte do miocárdio, intervenção coronária percutânea e cirurgia de revascularização do miocárdio. Como ponto de partida, a investigação teve por base a recolha dos dados de 528 pacientes de Paphos, no Chipre. Após a execução do processo de DM, os resultados apontaram para a idade, tabagismo e hipertensão como sendo os principais fatores de risco do enfarte do miocárdio. Por sua vez, quanto à intervenção coronária percutânea, os principais indicadores de risco foram os antecedentes familiares, a hipertensão e a diabetes. Já relativamente à cirurgia de revascularização do miocárdio, os atributos que aparentaram ser mais significativos para o desenvolvimento deste tipo de evento cardíaco foram a idade, a hipertensão e o tabagismo. Além disso, foram também apuradas as percentagens de acerto associadas a cada um dos eventos em estudo, registando-se uma acurácia de 66% para o enfarte do miocárdio e de 75% para a intervenção coronária percutânea e para a cirurgia de revascularização do miocárdio.

Posteriormente, Rajeswari, Vaithiyanathan e Neelakantan (2012) publicaram um artigo, no qual apresentaram uma proposta de seleção de atributos para a aplicação da técnica NN, com o objetivo de diminuir os custos e o tempo de processamento do algoritmo, e de aumentar a sua acurácia. Para tal, iniciaram o seu estudo através de um *dataset* relativo à doença cardiopatia isquémica, com 17 atributos e respeitante a um conjunto de 712 pacientes. Os autores testaram a técnica NN com todas as combinações possíveis entre atributos e propuseram a sua redução de 17 para 12, registando-se uma acurácia do modelo de 82.2% na fase de teste. Com base nesta proposta, concluíram que os atributos associados à idade, sexo, menopausa, IMC, circunferência da cintura, pressão arterial alta e baixa, diabetes, colesterol e tipo A foram os mais relevantes para a classificação do grau de risco de se contrair uma cardiopatia isquémica. Note-se que, neste caso, a designação "tipo A" se refere ao comportamento dos indivíduos que são agitados e muito ligados ao trabalho.

Mais tarde, Abdar *et al.* (2015) realizaram um estudo no sentido de apurarem as melhores técnicas de DM para a previsão de DCV. Para isso, utilizaram um conjunto de dados, proveniente da Universidade da Califórnia (Irvine), constituído por 13 campos e 270 registos. Os autores observaram que, entre as técnicas DT, NN, *Support Vector Machine* (SVM) e *K-Nearest Neighbors* (KNN), aquela que correspondeu aos melhores resultados foi a DT, que

registou, na fase de teste, valores de 93.02% para a acurácia e de 95.23% para a sensibilidade. Por outro lado, os valores de acurácia e de sensibilidade relativos à SVM foram de 86.05% e de 80.95% e os do algoritmo KNN foram de 88.37% e de 88.09%, respetivamente. No geral, a técnica NN foi a que se refletiu em piores resultados, com uma acurácia de 80.23% e uma sensibilidade de 73.80%. Este estudo mostrou ainda que, no caso da DT, os atributos mais relevantes para o contexto preditivo foram os dados clínicos relativos ao *thal* (talassemias), ao *slope* (a inclinação do pico ST em prova de esforço) e ao tipo de dor no peito.

Recentemente, Kim e Kang (2017) utilizaram a técnica NN para preverem também os riscos de se contraírem DCV, com recurso a um *dataset* que continha 4146 registos de indivíduos da Coreia do Sul. Como esta técnica tem um funcionamento do tipo "caixa-negra", em que as relações e as importâncias dos atributos não são visíveis, o autor propôs uma abordagem em duas etapas, de forma a evidenciá-las. A primeira fase serviu para se determinarem os atributos relevantes para a previsão e a segunda consistiu na construção de um modelo preditivo baseado em NN, considerando-se apenas os atributos selecionados. Deste modo, através desta metodologia, filtraram-se apenas os principais atributos que aumentam o risco de se sofrer de DCV, suprimindo-se os restantes.

De um modo mais detalhado, no primeiro processo analisou-se, individualmente e através da técnica NN, a sensibilidade associada a cada um dos atributos em estudo. De seguida, hierarquizaram-se esses atributos de acordo com o seu valor de sensibilidade e descartaram-se, gradualmente, os menos relevantes. De cada vez que se retirava um dado atributo, executava-se de novo a técnica e apurava-se se o desempenho do algoritmo tinha aumentado. Este processo repetiu-se para cada um dos atributos até se constatar uma pior capacidade preditiva do modelo, após o descarte de algum deles.

Por outro lado, na segunda parte da investigação, foram estudadas as correlações entre os atributos analisados pela técnica NN. Neste caso, sempre que um atributo fosse afetado na sua capacidade preditiva pela variação de outro, os autores consideravam-nos correlacionados.

Os resultados do primeiro processo mostraram que os atributos mais significativos para a previsão eram a idade, o IMC, o colesterol total, o colesterol HDL, a pressão arterial alta, a pressão arterial baixa, os triglicerídeos e a diabetes. Adicionalmente, fruto do segundo processo, identificaram-se os pares de atributos correlacionados IMC e pressão arterial baixa, colesterol total e pressão arterial baixa, e pressão arterial alta e pressão arterial baixa. Além disso, verificou-se também que os atributos idade, colesterol HDL, tabagismo e diabetes não se encontravam correlacionados com quaisquer outros parâmetros.

A metodologia adotada pelos autores seguiu a estrutura apresentada na **Figura 2**.

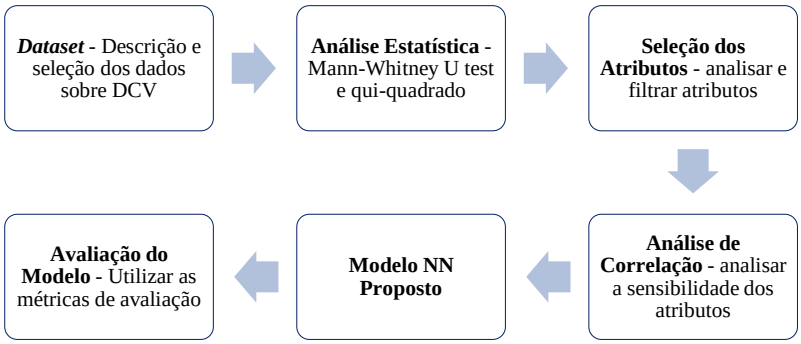


Figura 2 – Metodologia seguida por Kim e Kang (2017).

Numa análise final, os autores constataram que o modelo construído tinha a mesma capacidade preditiva que o FRS, na fase de treino, e que, na fase de validação, apresentava melhores valores da curva ROC, estando, assim, melhor adaptado à população coreana do que o FRS.

Todos estes estudos provam que as técnicas de DM se podem aplicar, com sucesso, ao contexto de predição de risco de DCV, ao permitirem priorizar os fatores de risco e criar modelos com a complexidade que uma investigação desta tipologia exige. Desta forma, estas técnicas podem ser utilizadas como suporte aos simuladores de risco de DCV atuais, funcionando como uma alternativa às técnicas de cálculo tradicionais existentes.

2.4 A Importância da Promoção de uma Monitorização Contínua

Além do algoritmo de cálculo, uma outra limitação dos simuladores atuais prende-se com o facto de se vocacionarem mais para uma medição pontual do risco de DCV, contabilizando apenas uma vez registos relativos a parâmetros que podem apresentar uma grande volatilidade diária, como é o caso dos valores das pressões arteriais.

Desta forma, os simuladores existentes não estão particularmente orientados para uma monitorização contínua, que permita guardar o histórico de cada utilizador e, com isso, avaliar as variações do valor de risco. Este controlo possibilitaria efetuar estimativas futuras mais rigorosas da evolução do risco e detetar a existência de medições atípicas, atenuando a sua influência no cálculo do risco geral. Assim, uma das principais desvantagens destes simuladores reside no facto de considerarem apenas um único registo como sendo representativo do estado clínico de cada utilizador, não sendo capaz de identificar valores pontualmente anormais, que só uma monitorização contínua conseguiria detetar. Em função das medições anteriormente efetuadas pelos utilizadores serem desprezadas, os resultados obtidos por estes simuladores são

totalmente independentes dos anteriores e o risco atual dos utilizadores não reflete os valores passados. Em consequência disso, os resultados obtidos são gerados de uma forma generalista e pouco minuciosa, em que a medição incorreta de um dos parâmetros pode comprometer todo o resultado final.

Uma hipótese plausível poderia passar pela recolha e pelo armazenamento dos dados fornecidos pelos utilizadores em *data warehouses* (DW). A implementação de repositórios deste tipo possibilitaria a incorporação de todos os registos históricos dos utilizadores, sem perda de dados, de uma forma organizada, com indicação temporal, consistente, rápida, adaptável a mudanças, facilmente acessível e com possibilidade de integração com outras ferramentas, como *dashboards* (Kimball and Ross, 2013).

A vantagem mais evidente da sua implementação seria o facto de as medições retornadas serem mais fiáveis. No entanto, além deste ponto, a integração dos dados num DW possibilitaria ainda a sua incorporação com ferramentas de visualização gráfica, que poderiam servir de incentivo aos utilizadores para a adoção de comportamentos que diminuíssem o seu valor de risco. Assim, para atingirem os valores pretendidos, os utilizadores seriam motivados a optar por um estilo de vida mais saudável e, com isso, melhorariam o seu bem-estar cardíaco.

Os simuladores atuais não apresentam também uma forte índole prática direcionada para o utilizador comum, na medida em que refletem valores probabilísticos de risco de contração de DCV, sem grande significado. Assim, dado que os resultados não proporcionam uma interpretação imediata, estes indicadores não fomentam a adesão de um grande número de utilizadores. Para que tal aconteça, os sistemas devem assegurar uma leitura intuitiva e uma fácil análise, capaz de refletir resultados que os utilizadores consigam percecionar de um modo intuitivo e imediato.

A conversão do valor probabilístico relacionado com o risco de DCV para um valor associado ao índice de bem-estar cardíaco seria uma das medidas possíveis para facilitar a compreensão dos resultados. Além disso, o uso complementar de *dashboards*, ou seja, de instrumentos de visualização gráfica, também poderia apresentar grande potencial para simplificar esta análise e ainda para servir de auxílio para fins estatísticos.

Em relação a estas ferramentas de observação gráfica, Stadler *et al.* (2016) descreveram o desenvolvimento de *dashboards* para uma unidade de saúde com o propósito de melhorar as visualizações dos dados dos pacientes. O estudo incidiu nas áreas da septicemia e das readmissões hospitalares em períodos inferiores a 30 dias. A construção destes *dashboards* permitiu observar estatísticas resumidas e destacar as métricas mais relevantes, além de fornecer informações visuais de fácil compreensão por parte dos pacientes e do pessoal

hospitalar, independentemente do seu nível de conhecimento e de especialização. Assim, a sua implementação possibilitou a identificação rápida de tendências e de oportunidades de melhoria e a sua utilização automatizada permitiu aumentar a eficiência e consolidar as métricas de desempenho, para que os diversos temas pudessem ser comparados com outras unidades de saúde. Os autores concluíram que, após a implementação dos sistemas de *dashboards*, existiu uma poupança efetiva anual de 289, 364 e 450 horas de trabalho, relativas ao primeiro, segundo e terceiro ano, respetivamente.

As mais valias dos *dashboards* foram ainda destacadas em mais estudos, como o de Hay *et al.* (2013). Durante o processo de investigação, estes autores abordaram a temática relativa à criação de mapas coloridos de toda uma área geográfica, com indicação visual dos locais onde as doenças infeciosas tinham sido identificadas. Nesses mapas, consoante a extensão da presença de infeções, foram associadas cores, de modo a apresentarem graficamente, e de uma forma simples, o panorama do estado da evolução das doenças infeciosas. Estes autores analisaram ainda as questões associadas com a elaboração de mapas de risco dinâmicos de doenças infeciosas, atualizados em tempo real a partir de registos existentes *online*. Nessa análise, consideraram que ainda havia uma grande falta de dados disponíveis, mas que a existência desses mapas daria uma extensão dos problemas relativos às infeções e auxiliaria a determinar a origem das suas causas.

3. A DESMISTIFICAÇÃO DOS PADRÕES OCULTOS

Os dados encerram em si padrões e relacionamentos ocultos que, perante os meios mais apropriados, podem ser desvendados e convertidos em conhecimento útil. Uma das formas de se revelar os seus segredos passa pela utilização de técnicas de DM, que permitem tornar visível a influência de fatores específicos num resultado final, além de possibilitarem também a avaliação do efeito da interação conjunta desses fatores. Nesse sentido, pretendeu-se criar um sistema capaz de retornar os índices associados ao bem-estar cardíaco de utilizadores comuns, de uma forma rigorosa, com recurso a técnicas de DM. Para tal, no presente capítulo foi seguida a metodologia *CRoss Industry Standard Process for Data Mining* (CRISP-DM), que pressupõe a execução de um conjunto de etapas sequenciais: a compreensão do negócio, a compreensão dos dados, a preparação dos dados, a modelação, a avaliação e, no final, a implementação.

3.1 Compreensão do Negócio

As DCV apresentam-se, atualmente, como uma das principais preocupações ao nível da Saúde nos países desenvolvidos. Estas doenças são uma das principais causas de morte nestes países e, em 2017, em Portugal, segundo dados do PORDATA, as doenças do aparelho circulatório foram responsáveis por 29.3% das mortes que ocorreram nesse ano. Além disso, de acordo com a Organização Mundial de Saúde (OMS), estima-se que 80% do desenvolvimento prematuro das DCV seja desencadeado por fatores de risco evitáveis, que poderiam ser combatidos por tratamento médico ou pela adoção de estilos de vida mais saudáveis. Apesar de existirem fatores de risco não evitáveis, como é o caso da idade, do sexo, da etnia e do histórico familiar, existem outros fatores que podem ser modificados pela alteração dos comportamentos ou pelo acompanhamento médico. Um dos modos de se facilitar a compreensão, por parte dos utilizadores, da influência dos fatores de risco no bem-estar cardíaco é através de índices. Os índices permitem converter toda a informação relevante do utilizador num único valor que, de acordo com uma escala pré-definida, reflete o seu grau de risco de desenvolvimento de uma DCV. Por norma, estes índices estão associados a uma escala de cores, o que facilita a apreensão da informação pela generalidade dos indivíduos. Neste âmbito, o objetivo principal do negócio passa pela criação de um mecanismo capaz de retornar índices de bem-estar cardíaco, que permitam aferir o grau de risco de desenvolvimento de uma DCV para um indivíduo em particular, dotando-o, assim, de um modelo indicativo do seu bem-estar. Deste modo, o sucesso do projeto prende-se com a disponibilização, aos utilizadores, de

uma ferramenta informativa do seu índice de risco cardíaco, que os permitirá orientar no sentido de reduzir os comportamentos individuais de risco.

Para se proceder à implementação do sistema em estudo foi fundamental dispor-se de um conjunto de recursos. Para tal, foi preciso assegurar-se a existência de ficheiros de dados classificados e de ferramentas apropriadas para o desenvolvimento do projeto de DM. Neste caso, o *software* utilizado foi o Pentaho Data Integration (Spoon) e o Pentaho Data Mining (Weka). Para se garantir a viabilidade do sistema foi preciso ainda satisfazer-se um conjunto de requisitos relacionados com o *software* e com os dados em análise. No que toca aos requisitos de *software*, foi requerida a instalação da versão profissional do Spoon, para que se pudessem executar os módulos relativos ao DM, que permitem a interligação com o Weka. Além disso, teve de se instalar a versão 3.7 do Weka (ou anterior), uma vez que as edições posteriores não permitem conectar o Spoon ao *workflow* do Weka. Por outro lado, os requisitos associados aos dados prendem-se com a necessidade de existirem dados contemporâneos e representativos da população atual.

Os índices criados devem refletir valores precisos sobre o grau de risco de desenvolvimento da doença. Para tal, estes índices devem ser determinados de uma forma rigorosa e adaptada de acordo com os valores de cada atributo. A escolha do algoritmo de cálculo reveste-se, por isso, de uma particular importância e os métodos tradicionais não permitem estimar índices com o rigor exigido a nível médico. Em contexto clínico de previsões médicas, os valores determinados devem ser exatos e, em caso de erro, é preferível que os índices informem sobre um cenário pessimista do que o contrário, para que se possam detetar todos os casos de doença e incentivar a ida ao médico de todos os possíveis doentes, de modo a prevenir-se e a detetar-se o surgimento da doença ainda num estágio precoce. Desta forma, as técnicas de DM funcionam como uma alternativa sólida aos métodos convencionais de cálculo, e têm como principal objetivo servir de base para a determinação dos índices, através da criação e da seleção dos algoritmos mais adequados, tendo em consideração a maximização da percentagem de acerto e a diminuição dos falsos negativos. Neste caso, estas técnicas de DM devem ser capazes de gerar modelos que consigam prever, a partir dos atributos em estudo, a probabilidade associada a cada registo de desenvolvimento de DCV. Assim, com base nesses valores probabilísticos, pode-se recorrer, numa segunda etapa, a funções matemáticas que permitam transformá-los, de um modo adequado, em índices. Desta forma, atingindo-se os objetivos do DM possibilita-se a concretização dos objetivos do negócio.

O modo de se avaliar a qualidade e o sucesso dos algoritmos de DM criados é através dos valores dados pela acurácia e pelos falsos negativos. Por um lado, o algoritmo escolhido deverá apresentar valores elevados de acerto e, em simultâneo, deverá ser capaz de evitar as situações de indivíduos com doença que, incorretamente, sejam identificados como sendo saudáveis. Assim, considerou-se como métrica de sucesso a existência de um modelo que apresentasse valores de acurácia e de sensibilidade superiores a 2/3, isto é, superiores a 67%.

3.2 Compreensão dos Dados

Em relação ao *dataset*, os dados selecionados para o processo de DM corresponderam à adaptação de registos provenientes de uma fonte de dados disponibilizada *online* na plataforma Kaggle, que contabilizou um total de 65000 instâncias e de 18 atributos. Nos 18 parâmetros considerados inclui-se o atributo de classificação (classe), que indica se um determinado registo é relativo a um indivíduo que sofre de DCV ou a um indivíduo que não tem a doença, e a data de captura dos registos. Além da data, os restantes atributos eram dos tipos numérico e nominal e, por isso, representavam variáveis quantitativas e qualitativas, respetivamente. Para a representação dos valores possíveis para os atributos nominais, a fonte de dados tem uma terminologia definida, em que o nível de exercício físico é dado por valores categóricos entre 0 e 3 e o sexo é representado por "0" e "1", para indivíduos do género feminino e masculino, respetivamente. Os restantes atributos nominais também podem assumir os valores "0" e "1", relativos às expressões "não" e "sim", por esta ordem. Além disso, todos estes parâmetros abrangiam os principais indicadores da literatura, relacionados com o desenvolvimento de DCV.

Para uma melhor compreensão dos dados, de forma a apurar-se a distribuição inicial dos valores de cada atributo do *dataset* em análise, elaboraram-se gráficos de barras a partir do *software* Weka, um para cada atributo distinto, que permitiram comparar as frequências com que os valores dos registos surgiam. Por uma questão de simplificação, as datas dos registos não foram incluídas nesta análise gráfica. Na **Figura 3** apresentam-se os resultados obtidos, sendo a cor azul representativa de indivíduos sem doença e a cor vermelha relativa aos que têm DCV.

Em relação à idade, constatou-se que os registos praticamente abrangiam indivíduos entre os 35 e os 64 anos, ou seja, o modelo de DM a ser construído será mais indicado para esse intervalo etário. Por outro lado, no que diz respeito ao sexo, 56.42% dos registos eram referentes a indivíduos do género feminino e 43.58% eram relativos ao género masculino. Assim, para

este parâmetro, observou-se uma proporção equilibrada de indivíduos inquiridos de ambos os sexos, representativa da população atual.

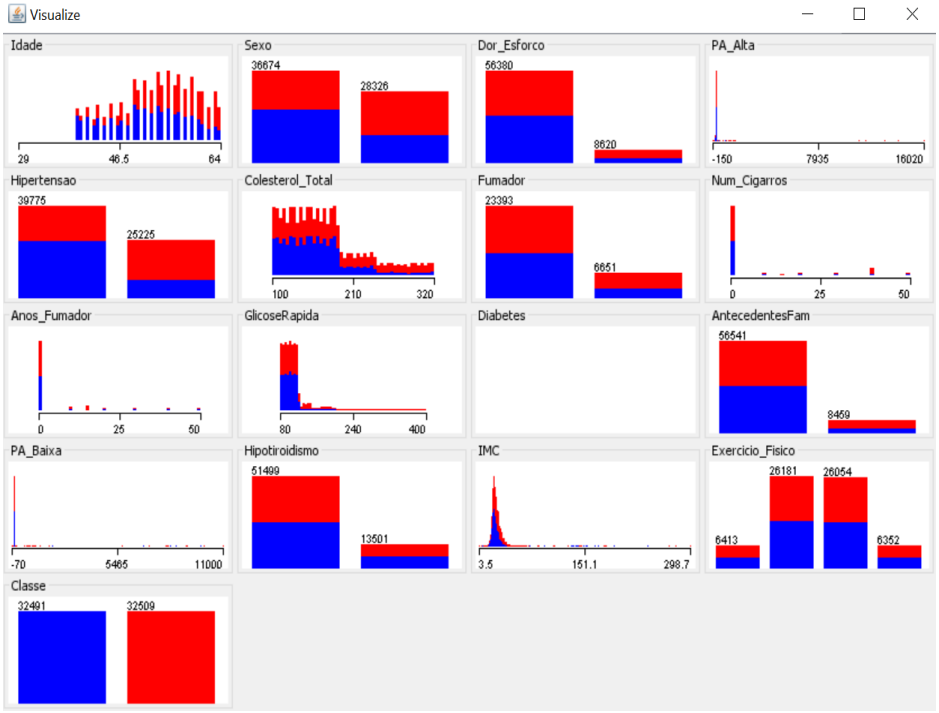


Figura 3 – Composição e distribuição inicial do *dataset* de DM.

Um dos parâmetros que se destaca nos resultados obtidos é o que diz respeito à diabetes, pelo facto de não existir nenhum registo com este valor preenchido. Nestas circunstâncias, para efeitos de análise, este atributo deverá ser excluído por não apresentar a disponibilidade necessária. No entanto, caso sejam efetuados novos processos de DM e o *dataset* passe a incluir estes valores preenchidos em número suficiente, o sistema deverá estar preparado para os ter em consideração nas novas análises.

Quanto aos atributos relativos à pressão alta, aos cigarros fumados, aos anos de fumador, aos níveis de glicose rápida, à pressão baixa e ao IMC, verificou-se a existência de valores extremos e, consequentemente, os seus histogramas não permitiram retirar conclusões acerca da distribuição destes dados. Nesse sentido, construiu-se a **Tabela 2**, com o propósito de se facilitar esta análise. Deste modo, a partir da tabela criada, concluiu-se que estes parâmetros apresentavam uma dispersão adequada e que poderiam ser considerados representativos.

Para os restantes gráficos gerados, verificou-se que os dados estavam também devidamente representados, incluindo-se o do atributo de classificação. No caso concreto deste parâmetro, uma distorção dos dados, mesmo que representativa da população, pode originar constrangimentos no modelo preditivo, pelo facto de uma das possíveis classificações ter consideravelmente menos registos do que a outra, durante a fase de treino do modelo. Assim,

relativamente a este atributo, um equilíbrio no número de registos associados a cada classe é fundamental para que o processo de DM tenha sucesso. Neste caso, uma vez que os dados destes registos estavam equilibrados (49.99% relativos a indivíduos saudáveis e 50.01% de indivíduos com DCV), não seria necessária, à partida, a adoção de abordagens que "forçassem" o seu equilíbrio como, por exemplo, reduzir-se o número de registos associados à classe maioritária (*undersampling*) ou duplicarem-se os registos da classe minoritária (*oversampling*).

Tabela 2 – Distribuição dos dados dos atributos com valores extremos de acordo com intervalos

Atributo	Valor	Nº Registos	% (Total)
P.A. Alta	< 125	37881	58.3
	125-150	22368	34.4
	> 150	4751	7.3
Nº de Cigarros	< 1	23377	77.8
	1-30	3199	10.6
	> 30	3468	11.6
Anos de Fumador	< 5	23381	77.8
	5-20	4070	13.5
	> 20	2593	8.7
Glicose Rápida	<100	26276	40.4
	100-130	28715	44.2
	> 130	10009	15.4
P.A. Baixa	< 75	12541	19.3
	75-90	46777	72.0
	> 90	5682	8.7
IMC	< 25	23933	36.8
	25-35	35292	54.3
	> 35	5775	8.9

De um modo global, tendo em consideração apenas os resultados da **Figura 3** e desprezando-se a correlação entre atributos, a incidência de DCV aparenta estar mais relacionada com o sexo, a idade, a perceção de dor após esforço, a hipertensão, o facto de se ser fumador e os antecedentes familiares. De facto, as DCV são mais evidenciadas em indivíduos do sexo masculino, séniores, pessoas que sintam dor após a prática de esforço, hipertensos, fumadores, ou indivíduos que tenham antecedentes familiares.

De forma a averiguarem-se as estatísticas básicas, como o número total de valores omissos e de valores distintos, foi construído, com recurso ao sistema R, o quadro que se apresenta na **Figura 4**, que resume esses resultados para cada um dos atributos, de acordo com o seu tipo. Além destes valores, no caso dos atributos inteiros, este resumo possibilita ainda uma análise à média, ao desvio-padrão e aos valores extremos de cada um deles.

```
> FonteDM = read.csv("C:/Users/Ana/Desktop/FonteDM.csv", header=TRUE, sep = ";")
> library('skimr')
> skim(FonteDM)
Skim summary statistics
 n obs: 65000
 n variables: 18

-- Variable type:factor -------------------------------------------------------
         variable missing complete     n n_unique                                top_counts ordered
  AntecedentesFam       0    65000 65000        2                 Nao: 56541, Sim: 8459, NA: 0   FALSE
           Classe       0    65000 65000        2                Doe: 32509, Sem: 32491, NA: 0   FALSE
      Dor_Esforco       0    65000 65000        2                 Nao: 56380, Sim: 8620, NA: 0   FALSE
 Exercicio_Fisico       0    65000 65000        4 Ele: 26181, Mod: 26054, Bai: 6413, Nen: 6352   FALSE
          Fumador       0    65000 65000        3          emp: 34956, F: 23393, V: 6651, NA: 0   FALSE
      Hipertensao       0    65000 65000        2                 Nao: 39775, Sim: 25225, NA: 0   FALSE
   Hipotiroidismo       0    65000 65000        2                 Nao: 51499, Sim: 13501, NA: 0   FALSE
 i..Data_Registo       0    65000 65000     3501           15/: 35, 18/: 34, 25/: 34, 27/: 34   FALSE
             Sexo       0    65000 65000        2                 Fem: 36674, Mas: 28326, NA: 0   FALSE

-- Variable type:integer ------------------------------------------------------
         variable missing complete     n   mean     sd   p0 p25 p50 p75  p100     hist
     Anos_Fumador   34956    30044 65000   5.43  12.01    0   0   0   0    50 ▇▁▁▁▁
 Colesterol_Total       0    65000 65000 170.14  52.48  100 130 160 190   320 ▇▇▅▂▁
    GlicoseRapida       0    65000 65000 119.58  56.43   80  92 104 116   400 ▇▁▁▁▁
            Idade       0    65000 65000  52.8    6.76   29  48  53  58    64 ▁▂▃▆▇
     Num_Cigarros   34956    30044 65000   7.08  14.57    0   0   0   0    50 ▇▁▁▁▁
          PA_Alta       0    65000 65000 128.94 159.55 -150 120 120 140 16020 ▇▁▁▁▁
         PA_Baixa       0    65000 65000  96.61 188.06  -70  80  80  90 11000 ▇▁▁▁▁

-- Variable type:logical ------------------------------------------------------
 variable missing complete     n mean count
 Diabetes   65000        0 65000  NaN 65000

-- Variable type:numeric ------------------------------------------------------
 variable missing complete     n  mean   sd  p0  p25  p50  p75  p100     hist
      IMC       0    65000 65000 27.57 6.12 3.5 23.9 26.4 30.2 298.7 ▇▁▁▁▁
```

Figura 4 - Estatísticas básicas relativas a cada um dos atributos do *dataset*.

Da visualização do quadro destacam-se alguns valores desajustados, que poderão ter correspondido a erros de inserção dos dados no *dataset*. A título de exemplo, verifica-se que os valores mínimos e máximos das pressões arteriais não podem corresponder à realidade, na medida em que os primeiros assumem valores negativos, ao passo que, os valores máximos são da ordem dos milhares. Além disso, no que diz respeito ao IMC, os seus valores extremos também não são razoáveis, uma vez que o valor mínimo registado é anormalmente baixo (3.5), enquanto o máximo é demasiado elevado (298.7). Assim, facilmente se depreende que estes valores não são aceitáveis e que, para o processo de DM decorrer com sucesso, eles não devem ser considerados. Deste modo, para que a capacidade preditiva dos algoritmos não seja negativamente influenciada, devem-se assegurar intervalos de valores coincidentes com a realidade esperada para cada atributo.

O quadro em estudo mostra ainda que, à semelhança do que já se tinha constatado, a diabetes não apresenta campos preenchidos, estando a totalidade dos registos associada a valores omissos. Além deste atributo, apenas o número de anos em que os indivíduos já fumaram ou continuam a fumar e o número de cigarros fumados, em média, por dia, é que apresentam valores omissos.

Além desta observação quantitativa, foi também efetuada uma apreciação visual, com o objetivo de se perceber o relacionamento existente entre os atributos. Para isso, cada um dos atributos foi comparado graficamente com outro, numa análise entre pares, de modo a verificar-se a existência de situações de linearidade entre eles. Quando existentes, estas situações de

linearidade evidenciam casos de atributos correlacionados, na medida em que uma variação num dos atributos, conduz a variações nos valores do outro. Após a avaliação dos relacionamentos entre todos os pares de atributos, apenas se observou uma possível relação entre os atributos relativos às pressões arteriais alta e baixa, que é visível na **Figura 5**. Como se pode constatar da visualização desta figura, os pontos, na parte inferior do gráfico, sugerem uma reta inclinada, que pressupõem algum nível de dependência. No entanto, como as pressões arteriais alta e baixa apresentam valores extremos que não são compatíveis com a realidade, a escala do gráfico pode não conduzir às interpretações mais adequadas.

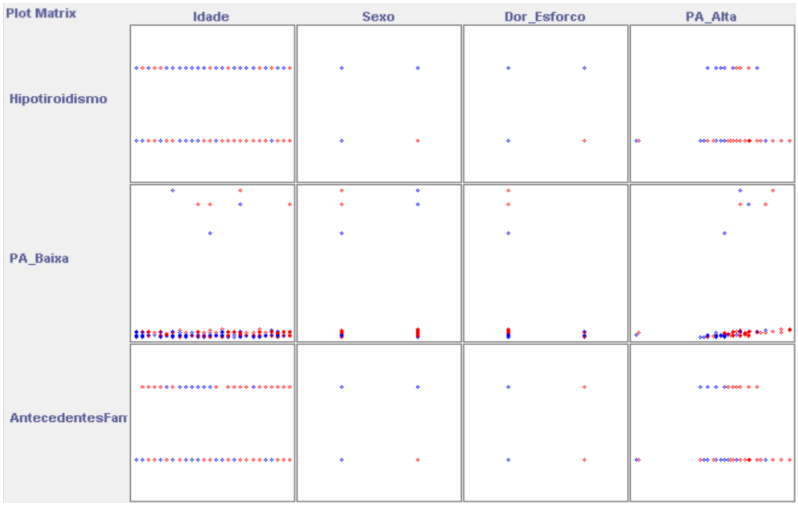


Figura 5 - Extrato da visualização gráfica entre pares de atributos.

Assim, para se comprovar a existência (ou não) de uma inter-relação entre este par de atributos, foi efetuada a ampliação ao gráfico ilustrada na **Figura 6**, que tem em conta apenas os valores aceitáveis para estes parâmetros. Contrariamente ao gráfico anterior, a partir deste não se conseguiu realçar um comportamento linear entre os atributos em estudo e, por isso, nenhum par de atributos do *dataset* revelou, de forma visível, uma correlação.

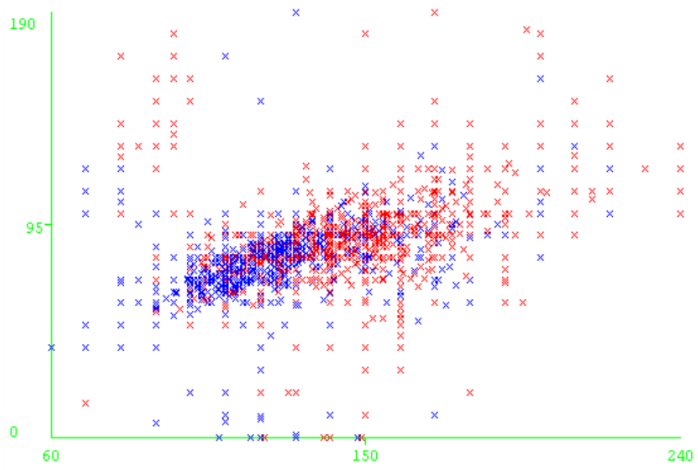


Figura 6 - Relacionamento da pressão arterial alta (eixo horizontal) com a pressão arterial baixa (eixo vertical).

3.3 Preparação dos Dados

Para se proceder à seleção do melhor algoritmo de DM, é necessário, numa primeira fase, proceder-se à limpeza e ao tratamento dos dados, de forma a corrigirem-se as deficiências encontradas. Esta etapa de preparação dos dados pode ser efetuada de uma forma automatizada, sem ser precisa uma prévia aprovação por parte do analista para que possa ser validada. Deste modo, os dados podem ser limpos e transformados automaticamente, sem necessidade de intervenção humana, de cada vez que seja desencadeado um processo de DM. No entanto, este procedimento não dispensa uma manutenção periódica para se poder garantir a sua viabilidade e os registos excluídos da análise devem ser guardados na forma de um ficheiro, para permitir que, posteriormente, o analista os possa observar e verificar se podem ser corrigidos e considerados em análises futuras.

A etapa de preparação dos dados foi executada, em duas etapas complementares, com recurso a *software* da Pentaho. Na primeira fase, utilizou-se a ferramenta gráfica Spoon e a segunda etapa foi realizada no Weka.

Na **Figura 7** está representada, de forma esquemática, a primeira fase de preparação dos dados para o processo de DM.

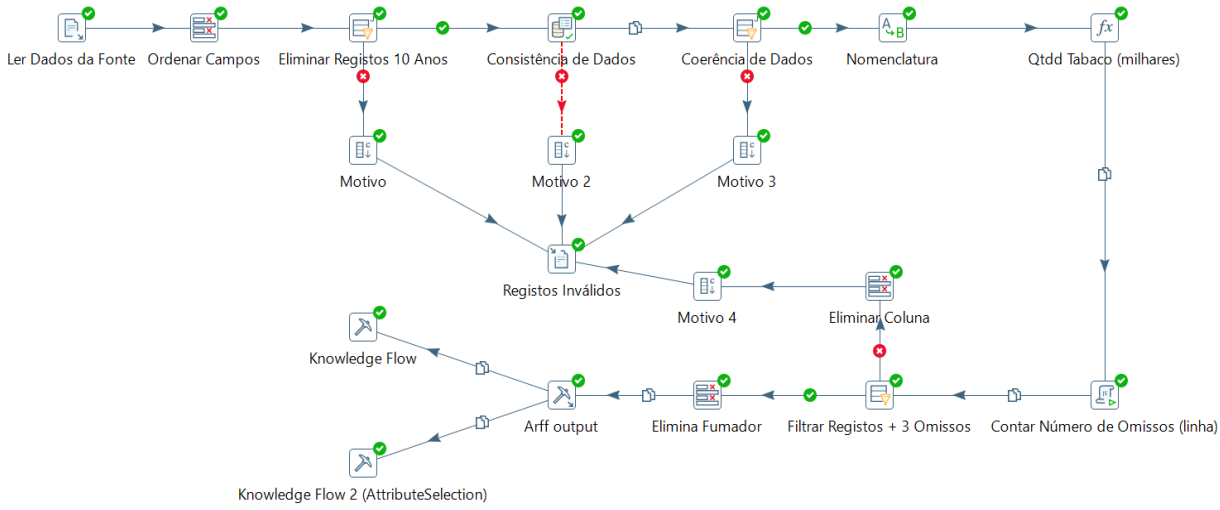


Figura 7 - Arquitetura do tratamento de dados para o processo de DM (Etapa 1).

Nesta fase, em primeiro lugar, foram extraídos os dados da fonte e ordenados todos os campos. A etapa seguinte consistiu na filtragem dos registos, em que se excluíram:

- Registos com dez anos ou mais.
- Dados inconsistentes.

- Dados incoerentes.

Tal como referido na secção anterior, os algoritmos utilizados para o cálculo do valor dos índices devem ter em conta o estilo de vida contemporâneo da população. Deste modo, os registos usados como base para o processo de DM devem aproximar-se temporalmente da data em que se determinam os índices. Nesse sentido, considerou-se que entre o período de cálculo dos índices e a data dos registos considerados para DM não deve existir um desfasamento superior a dez anos, para que a condição da contemporaneidade dos dados permaneça válida. Assim, todos os registos com mais de dez anos em relação à data em que os índices são determinados não devem ser considerados no processo de DM e, para isso, utilizou-se a condição *(org.apache.commons.lang.time.DateUtils.addYears(new java.util.Date(), -10).compareTo(Data_Registo)) < 0* para se filtrarem apenas os registos mais recentes.

Além disso, outro ponto importante passou por se garantir a consistência de todos os dados da fonte. Para tal, foi preciso limitar cada atributo quanto ao seu tipo de dados (como *string*, data ou inteiro) e quanto aos seus valores admissíveis. A título de exemplo, considerou-se que a idade apenas pode assumir valores compreendidos entre 18 e 130 anos e, por outro lado, o sexo, como se trata de uma variável categórica, apenas pode assumir os valores 0 e 1 (representativos de indivíduos do sexo feminino e masculino, respetivamente). Além disso, foi necessário identificarem-se ainda os atributos que não admitem valores nulos. Neste caso, o único parâmetro de cada registo que se considerou ter necessariamente de estar preenchido é o que diz respeito à sua data.

Ao nível da coerência, foi preciso eliminarem-se os registos que indicam que os indivíduos começaram a fumar antes dos dez anos de idade e registos contraditórios, que classifiquem os indivíduos como "fumadores", mas onde esteja indicado que o número de cigarros por eles fumado seja 0. Para isso, recorreu-se às condições indicadas na **Figura 8**.

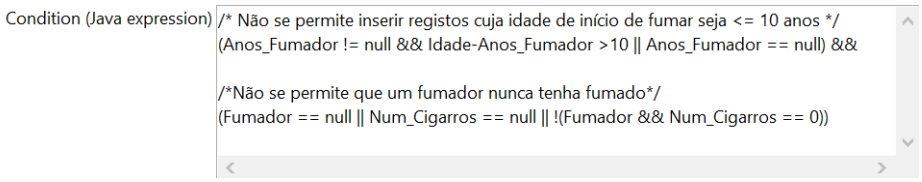


Figura 8 - Condições, em Java, para assegurar a coerência dos dados referentes ao tabagismo.

Após a filtragem dos registos, procedeu-se à integração das colunas relativas aos hábitos tabágicos num único atributo, através da fórmula indicada na **Equação 1**. Assim, inseriu-se uma nova variável, que correspondia ao número total de cigarros fumados, em milhares e, de

seguida, eliminaram-se as restantes colunas que diziam respeito aos fumadores, por serem redundantes com este novo atributo.

$$Qtdd\_Fumador = \frac{Anos\_Fumador \times Num\_Cigarros \times 365}{1000} \quad (1)$$

No final desta fase, para se impedir que os registos com mais do que três campos omissos fossem tidos em conta na análise de DM, foi preciso utilizar-se o código, em Java, da **Figura 9**, para se contar, inicialmente, o número de campos nulos de cada registo. Este código consistiu em "percorrer" todos os atributos de cada registo e, no caso de serem do tipo *null*, incrementou-se a variável associada aos valores omissos em uma unidade. Deste modo, obtendo-se uma nova coluna com o número de campos nulos de cada registo, foi possível filtrarem-se apenas os registos em que esse número era inferior a "3".

```
var fields = getInputRowMeta().getFieldNames();

var num_nulos = 0;
for (var i = 0; i < fields.length; i++) {
    if (row[i] == null) {
        num_nulos += 1;
    }
}
```

Figura 9 - Código Java para contagem de campos omissos por registo.

Depois de pré-processados os dados no Spoon, foi preciso guardá-los na forma de um ficheiro "arff", para possibilitar a sua posterior utilização no *software* Weka. Ainda nesta fase, foi preciso especificar-se, para cada atributo, o seu tipo, bem como identificar-se o atributo de classificação. Além disso, utilizou-se o *step Knowledge Flow*, de forma a permitir-se a comunicação automática do Spoon com o *software* Weka. Neste caso em específico, foram utilizados dois *steps Knowledge* Flow, de modo a invocarem os dois cenários distintos criados na fase da *Modelação*, que serão detalhados nesse capítulo.

Assim, a segunda etapa da preparação dos dados foi executada através do Weka e está esquematicamente representada na **Figura 10**.

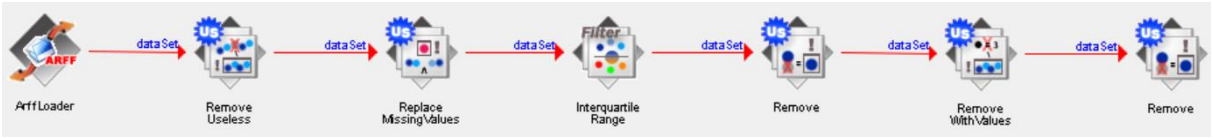


Figura 10 - Arquitetura do tratamento de dados para o processo de DM (Etapa 2).

Tendo-se como *input* o ficheiro em formato "arff" resultante do tratamento dos dados efetuado na primeira fase, no Spoon, foram utilizados vários filtros no Weka para se complementar essa etapa, de forma a estarem devidamente preparados para serem usados como base para o processo de avaliação das técnicas de DM.

Assim, para finalizar o tratamento dos dados, descartaram-se do modelo os atributos que tinham uma variância excessiva ou que, em contrapartida, praticamente tomavam sempre os mesmos valores em todos os registos.

Além disso, para se solucionar o problema dos campos em falta, optou-se pela sua substituição pela média ou pela moda do atributo a que diziam respeito, caso se tratassem de valores numéricos ou nominais, respetivamente. Outra alternativa, pela qual não se optou por poder implicar a perda de dados importantes para a previsão, seria a de se eliminarem todos os registos com valores em falta. Como anteriormente já se tinham filtrado apenas os registos com três ou menos campos omissos, considerou-se que todos aqueles que respeitassem esta condição deveriam ser incluídos para a construção do modelo de DM, de modo a evitar-se uma redução significativa do *dataset*. Uma opção adicional seria a de se retirarem todos os parâmetros que contivessem valores nulos, embora esta solução não fosse a mais indicada, uma vez que poderia implicar a exclusão de vários atributos que influenciassem substancialmente o resultado e, caso existisse de facto um grande valor de campos omissos para um dado atributo, essa coluna teria sido removida pelo filtro anterior, dada a sua baixa variância.

Outro aspeto importante a ter em conta é a existência de valores atípicos (*outliers*) e extremos que se possam detetar em campos numéricos. Os registos que contenham este tipo de valores devem ser removidos para não influenciarem negativamente a execução dos algoritmos de DM e, com isso, comprometerem o resultado do modelo. Para tal, o *software* Weka dispõe de mecanismos que permitem a sua identificação e remoção. Assim, com o intuito de se avaliar a presença de *outliers* e de valores extremos contidos no *dataset* em estudo, foram utilizadas a **Equação 2** e a **Equação 3**, respetivamente. Deste modo, todos os registos que continham, pelo menos, um valor situado fora dos intervalos dados por estas equações foram eliminados. De notar ainda que, neste caso, a sigla IQ se referia ao intervalo interquartil e FO e FVE aos fatores relativos aos *outliers* e aos valores extremos, respetivamente. No caso do presente trabalho foi considerado o valor "3" para FO e o valor "6" para FVE.

$$x > Quartil1 - FO \times IQ \wedge x < Quartil3 + FO \times IQ \quad (2)$$

$$x > Quartil1 - FVE \times IQ \wedge x < Quartil3 + FVE \times IQ \quad (3)$$

Após o pré-processamento dos dados, seguiu-se a fase de modelação do problema para a execução do DM.

3.4 Modelação

A etapa da Modelação é um passo determinante para o cálculo dos índices, uma vez que é nesta fase que se escolhem e avaliam os atributos mais relevantes e as técnicas mais adequadas a aplicar ao modelo. No caso do presente estudo, os índices exigem um processo de avaliação ainda mais incisivo, já que se tratam de valores relacionados com a Saúde e, nesse sentido, devem funcionar como uma ferramenta de apoio capaz de detetar todos os indivíduos prováveis de virem a sofrer de DCV ou convalescentes, e instigar a sua ida ao médico, priorizando mais o lado da prevenção da doença do que o seu tratamento. Deste modo, esta etapa requer uma avaliação minuciosa e um rigor acrescido para que os parâmetros e a metodologia selecionados para o cálculo reflitam não só bons valores de acerto, como também percentagens reduzidas de falsos negativos e tempos de execução baixos.

Em primeiro lugar, foram criados dois cenários distintos para o modelo considerado, de forma a apurar-se o conjunto de atributos que se repercute nos melhores resultados. Para tal, o primeiro cenário criado foi constituído por todos os atributos resultantes do tratamento dos dados efetuado e, por sua vez, o segundo cenário foi composto pela filtragem desses atributos com o *Attribute Selection*, do Weka. Este filtro foi configurado utilizando-se o avaliador de atributos *CfsSubsetEval*, que considera a redundância entre os atributos, e o método de pesquisa *Best-first*, que retorna a lista dos atributos que apresentam, localmente, a maior capacidade preditiva (Witten *et al.*, 2017). Em processos de DM, a redução dos atributos em análise poderá apresentar vantagens, como a diminuição do tempo de execução dos algoritmos, ou traduzir-se em melhorias ao nível da capacidade preditiva, ao eliminar atributos menos relevantes e/ou que tenham algum grau de dependência entre si. Cada um dos cenários criados corresponde a um processo de modelação distinto e, por isso, optou-se por se criar cada um deles num ficheiro diferente. Tendo-se como ponto de partida os dados considerados no *dataset* em análise, a constituição de cada um dos cenários é a que se descreve de seguida.

Cenário I: Idade, Sexo, Dor Após Esforço, Pressão Arterial Alta, Pressão Arterial Baixa, Hipertensão, Colesterol Total, Quantidade Fumador, Glicose Rápida, Antecedentes Familiares, Hipotiroidismo, IMC, Exercício Físico e Classe

Cenário II: Idade, Sexo, Dor Após Esforço, Pressão Arterial Alta, Pressão Arterial Baixa, Hipertensão, Colesterol Total, Glicose Rápida, Antecedentes Familiares, Exercício Físico e Classe

Note-se que, caso o *dataset* sofresse alterações, os atributos que comporiam estes cenários poderiam não ser os mesmos, uma vez que se trata de uma seleção automática, que tem por base os atributos resultantes do pré-processamento dos dados, que é também realizado automaticamente. A título de exemplo, o *dataset* inicial era também composto pelo atributo associado à presença de diabetes mas, após a etapa de tratamento de dados, este campo foi excluído por não ter um número suficiente de registos e, por conseguinte, não apresentar variabilidade nos dados.

Para a seleção das técnicas, foi efetuada uma recolha dos algoritmos mais comummente utilizados para a previsão de DCV. Apesar de não existir uma predominância nos resultados que possibilite uma avaliação hierárquica das técnicas, os trabalhos desenvolvidos neste âmbito recorrem frequentemente aos mesmos algoritmos de modelação, que já se mostraram eficientes na resolução deste tipo de problemas, como é o caso do *J48*, *Random Forest* (RF), NB, KNN, *MultiLayer Perceptron* (MLP) ou SVM. A título exemplificativo, a **Tabela 3** reúne as técnicas utilizadas em alguns estudos realizados neste contexto.

Tabela 3 - Exemplos de estudos desenvolvidos para previsão de DCV

Autores	Técnicas Utilizadas
(S.Dangare and S. Apte, 2012)	• MLP • J48 • NB
(Taneja, 2013)	• J48 • MLP • NB
(Bahrami and Hosseini Shirvani, 2015)	• J48 • KNN • NB • SVM
(Sreejith, Rahul and Jisha, 2016)	• KNN • J48 • NB • RF

Assim, a escolha das técnicas a utilizar no processo de DM, para o presente caso de estudo, foi baseada na literatura existente e recaiu sobre o conjunto de técnicas referido na **Tabela 3**, à exceção do algoritmo SVM, que não foi incluído na análise. A razão por não se ter considerado esta técnica como "candidata" ao cálculo dos índices residiu no facto de exigir um enorme esforço computacional e, por conseguinte, não conseguir lidar de forma eficiente com grandes conjuntos de dados, e também pelo facto de requerer a otimização de inúmeros parâmetros (Kecman, 2005; Pillai, 2017). Desta forma, os algoritmos selecionados foram o J48, RF, NB, KNN e MLP.

De uma forma sucinta, o algoritmo J48 cria uma árvore de decisão, na qual o nó superior representa o atributo com maior ganho de capacidade preditiva, ou seja, corresponde ao atributo mais relevante no processo de classificação de cada instância. Por sua vez, os nós subsequentes são construídos de um modo análogo, até todos os atributos terem sido considerados e, no final, obtém-se uma árvore com atributos hierarquizados. Cada caminho diferente da árvore corresponde a uma regra, que tem associada uma percentagem de acerto, relativamente à sua capacidade preditiva. (Jankowski and Jackowski, 2014; Witten *et al.*, 2017) No caso do presente trabalho, este algoritmo tem como vantagens a sua rapidez face a grandes volumes de dados e a produção de resultados facilmente entendíveis, que se podem transformar em regras (Bahrami and Hosseini Shirvani, 2015). Assim, no caso particular de previsões de DCV, este algoritmo permite averiguar, em termos gerais, quais são os atributos que mais influenciam o valor do índice e, através das regras, possibilita prever, para cada registo, o valor do índice mais provável que lhe está associado. Uma das principais limitações desta técnica é a possível inclusão de nós pouco relevantes, que aumentem a complexidade da árvore e, consequentemente, possam originar *overfitting*, ou seja, que "treinem" o modelo apenas para os dados em estudo. Deste modo, o J48 requer que os resultados obtidos sejam validados com novos dados de teste, distintos dos utilizados no processo de treino do modelo, para que se garanta que os valores de acurácia não são inflacionados pela existência de *overfitting*. (Jadhav and Channe, 2016; Joseph, Hlomani and Letsholo, 2016; Saravana and Gayathri, 2018)

Por sua vez, o RF pressupõe a implementação de um grande número de árvores de decisão, construídas de forma aleatória, sem se considerar a importância dos atributos na hierarquia dessas árvores. No final, o valor preditivo para cada registo será dado pelo valor da classe mais indicada, ao percorrerem-se todas as árvores de decisão geradas. (Han, Kamber and Pei, 2012) Como este algoritmo é utilizado em *datasets* de grandes dimensões e, normalmente, apresenta bons resultados, torna-se indicado para o problema em estudo. No entanto, é preciso ter em conta que este método apresenta uma sensibilidade elevada à presença de *outliers*,

embora no presente trabalho esta restrição não seja um problema, uma vez que todos os valores atípicos e extremos são removidos na fase do tratamento dos dados. Além disso, tem também a desvantagem de proporcionar uma interpretação mais difícil que a da técnica anterior, apesar das suas semelhanças entre elas. (Pillai, 2017)

Relativamente ao NB, este algoritmo efetua as previsões através do teorema de Bayes da probabilidade condicionada em cada atributo, apresentando como principais características a sua simplicidade e rapidez (S.Dangare and S. Apte, 2012). Como este algoritmo não utiliza funções iterativas, apresenta tempos de processamento muito reduzidos quando comparados com as restantes técnicas de DM, sobretudo em situações em que seja utilizado um grande volume de registos. Apesar da sua rapidez, o NB apresenta um bom desempenho, identificando e eliminando os atributos menos relevantes para a previsão. Por outro lado, uma das suas limitações é a assunção da independência entre os atributos e outra das suas desvantagens deve-se ao facto de requerer muitos registos para que se consigam obter bons resultados. (Jadhav and Channe, 2016; Joseph, Hlomani and Letsholo, 2016)

Quanto ao KNN, o modo de funcionamento deste algoritmo prende-se com a estimativa dos *k* vizinhos mais próximos para um dado registo e com a atribuição da classe maioritária desses *k* vizinhos. Esta técnica apresenta como principais vantagens o facto de ser de simples implementação e de ser pouco sensível ao ruído. Contudo, é preciso ter em conta as suas restrições, que se devem, essencialmente, a tempos de execução elevados, sobretudo em grandes conjuntos de dados, como é o caso do *dataset* a ser analisado, e a uma sensibilidade à estrutura local dos dados, que se pode traduzir em otimizações locais, em vez de globais. (Han, Kamber and Pei, 2012; Jadhav and Channe, 2016)

No que diz respeito ao MLP, esta técnica representa uma rede neuronal, que simula o comportamento dos neurónios do cérebro. Neste algoritmo existem, pelo menos, três níveis de conjuntos de neurónios: o nível de *input*, uma ou mais camadas escondidas e o nível de *output*. De uma forma simplificada, nesta técnica, cada um dos neurónios dos vários níveis recebe e transforma sinais de entrada em sinais de saída, através de funções de ativação, geralmente não-lineares. (Ho and Li, 2016) No âmbito do estudo desenvolvido, esta técnica aparenta ser adequada, por ser indicada para lidar com sistemas complexos. No entanto, é importante ter em consideração o seu tempo de execução, que normalmente é elevado na presença de *datasets* de grandes dimensões (Pillai, 2017).

Para se proceder à modelação do caso em estudo, tendo em conta as técnicas e os atributos selecionados, foi preciso planear a forma de se testar a qualidade e a validade dos modelos. Para tal, optou-se por se considerar a existência de dados de treino, de validação e de teste. Os

dados de treino e de validação dizem respeito à totalidade dos dados contidos no *dataset* em estudo, enquanto os dados utilizados para teste foram obtidos através de outro *dataset*, que continha 5.000 registos, com as mesmas colunas. De notar que ambos os *datasets* têm de ser sujeitos, inicialmente, à etapa de preparação dos dados já mencionada, para estarem devidamente tratados para serem analisados. Através do conjunto de dados de treino e de validação são criados os modelos, cada um deles correspondente à utilização de uma técnica distinta. Na fase de aprendizagem dos modelos, de forma a apurarem-se valores de acurácia e de falsos negativos, são utilizados também os dados de treino e de validação que, neste caso, têm de ser previamente divididos, para poderem funcionar como dados de treino e de teste. Para a partição dos dados em treino e validação, recorreu-se à utilização de *cross-validation*, tendo-se optado por uma configuração a 10 *folds*, uma vez que, segundo Witten *et. al.* (2017), é o número de *folds* convencionado como *standard* e o que melhor resultados proporciona de uma forma geral. Este método, de um modo sucinto, divide os dados em dez subconjuntos, utilizando nove desses subconjuntos para treino e um para teste, permitindo obter o valor da acurácia inicial dada pela utilização do modelo criado nos dados de teste. Após esta avaliação, altera-se o subconjunto escolhido para teste e este procedimento é efetuado repetidamente até que todos os dez subconjuntos tenham sido utilizados como teste, sendo a acurácia do modelo o valor correspondente à média das acurácias encontradas. No final, devem ser registados os valores de acurácia e de falsos negativos referentes a cada uma das técnicas e selecionado o modelo que melhor se adequa ao caso em estudo, tendo em conta estes parâmetros. Além disso, é preciso ainda avaliar-se se o modelo escolhido apresenta também resultados semelhantes quando submetido a diferentes dados de teste. Para tal, os 5.000 dados do segundo *dataset* permitem efetuar esta avaliação e apurar casos de existência de *overfitting* nos modelos. Na **Figura 11** ilustra-se, de uma forma esquemática, o planeamento efetuado para o problema.

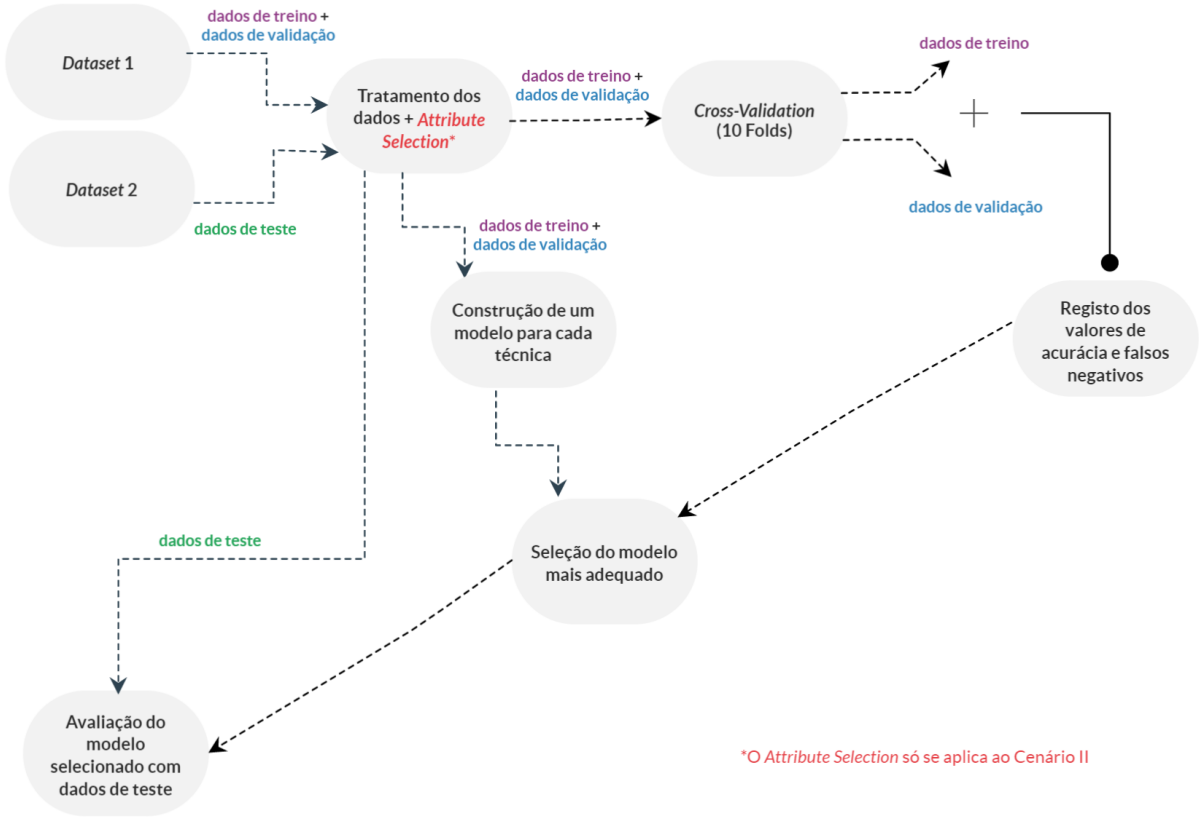


Figura 11 - Planeamento da metodologia a seguir para testar a qualidade e a validade do modelo.

À semelhança da *Preparação dos Dados*, a fase da *Modelação* foi implementada também com recurso aos *softwares* Weka e Spoon. Nesse sentido, a construção dos modelos e o registo dos valores de acurácia e dos falsos negativos foram realizados no Weka, ao passo que o processo de seleção do modelo mais adequado foi concretizado no Spoon. Para a sua posterior avaliação, através de dados de teste, utilizou-se, novamente, o Weka.

No Weka, tendo-se como parâmetro de entrada o *dataset* que continha os dados de treino e de validação já tratados, foi, numa primeira fase, definido o atributo de classificação (neste caso, o que se designa por "Classe") e, posteriormente, implementado um mecanismo de construção e de avaliação dos modelos. Na **Figura 12** está representado o procedimento efetuado no Weka, para o cenário I (no cenário II, a arquitetura é idêntica, diferindo apenas no facto de se usar o filtro "*Attribute Selection*" após ser definido o atributo de classificação). Na parte superior da figura, é efetuada a avaliação de cada uma das técnicas e as métricas resultantes são guardadas em ficheiros de texto, em que a designação de cada ficheiro corresponde ao nome da técnica utilizada, de forma a facilitar a sua identificação. Por sua vez, na parte inferior está esquematizado o processo de construção de cada um dos modelos. Também nesta etapa, os modelos são guardados, atribuindo-se o nome da técnica utilizada ao

nome do ficheiro do modelo. É de notar ainda que, para se distinguirem os nomes dos ficheiros relativos às métricas e aos modelos resultantes dos cenários I e II, se atribuiu, adicionalmente, o prefixo "*AS*" ao nome de todos os ficheiros gerados com base no cenário II.

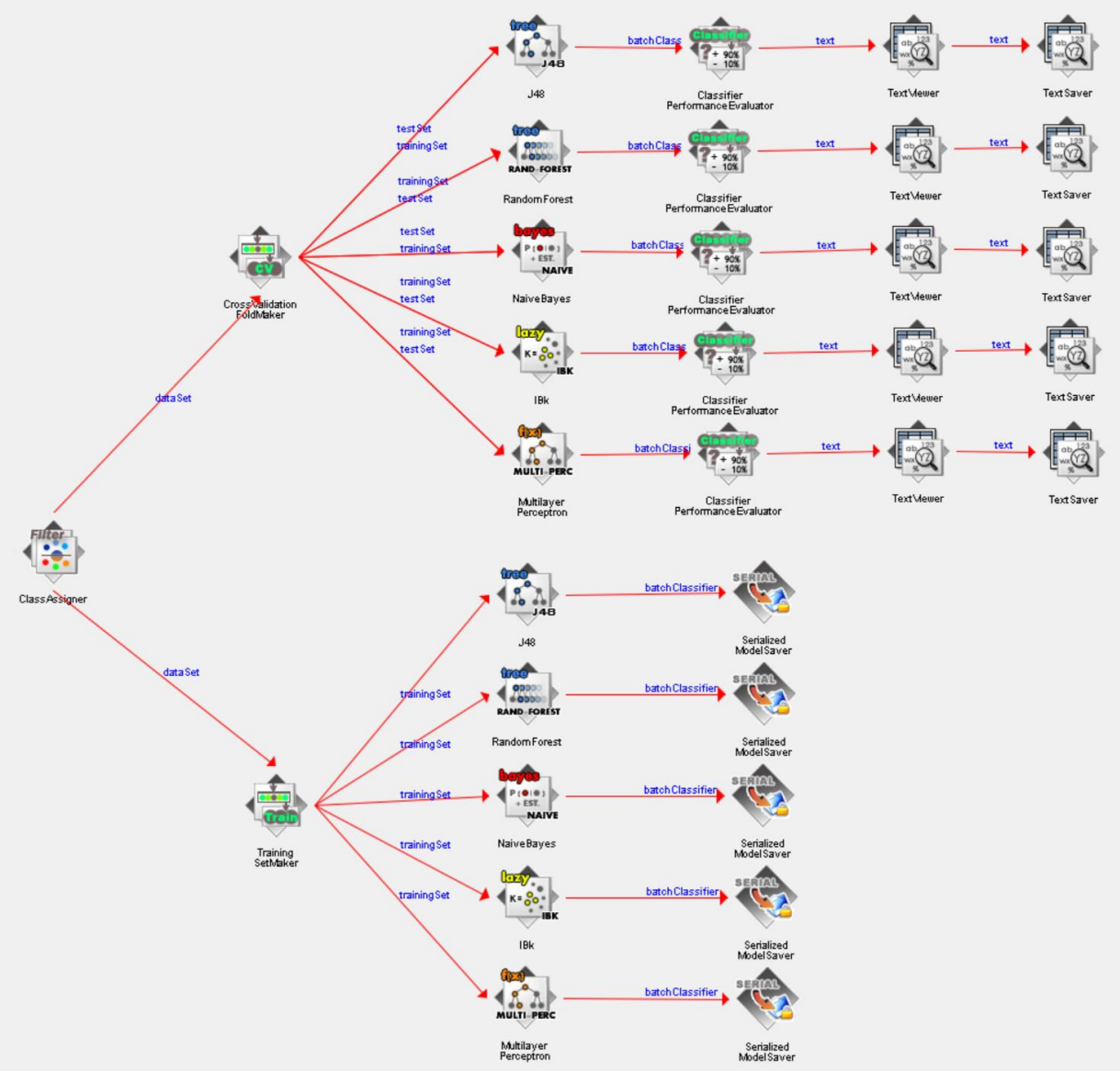


Figura 12 - Arquitetura da modelação do processo de DM.

Ainda antes de se proceder à construção dos modelos e à sua avaliação, foi preciso configurar-se os parâmetros de cada uma das técnicas utilizadas, com o intuito de se averiguar quais os seus valores mais adequados para a maximização da acurácia e para a minimização dos falsos negativos. Além disso, foi importante garantir-se que a escolha destes parâmetros se refletia em tempos de execução aceitáveis relativos quer à construção dos modelos, quer à sua avaliação.

O algoritmo J48 é composto pelos campos *unpruned*, *reduced error pruning* e pelo fator de correção MDL, que podem estar ativos ou inativos, e ainda por um fator de confiança, um

número mínimo de objetos e um número de *folds* associados. Para o processo de *tuning*, fizeram-se variar estes parâmetros de acordo com os valores indicados na **Tabela 4**. Uma prática comum neste algoritmo, que se designa por *pruning*, consiste na redução dos nós menos significativos da árvore gerada o que, consequentemente, pode resultar numa diminuição do tempo de execução e do risco de existência de *overfitting*. Por princípio, desde que não sejam detetadas diminuições bruscas no valor da acurácia, a sua utilização deverá ser preferível. (S.Dangare and S. Apte, 2012) O campo *unpruned* do J48 relaciona-se inversamente com esta prática, na medida em que o seu valor "falso" corresponde a uma ativação do *pruning*. Assim, sempre que este parâmetro assumir o valor "falso", as árvores geradas serão mais simples e menos propensas a situações de *overfitting*. Em contrapartida, a adoção deste procedimento poderá implicar perdas na capacidade preditiva e, por isso, o seu uso deverá ser ponderado no caso de se gerarem quebras significativas no valor da acurácia (Patel and Upadhyay, 2012). Por outro lado, o campo *reduced error pruning* é uma forma de *post-pruning*, uma vez que executa *pruning* após a construção da árvore, utilizando a minimização dos erros para o corte (Quinlan, 1993; Patel and Upadhyay, 2012). Deste modo, caso este campo esteja ativo, tem de existir, necessariamente, *pruning* (ou seja, *unpruned* terá de ser "falso"). Já o fator de correção MDL (*Minimum Description Length*) é um parâmetro que é responsável, quando se encontra ativo, por encontrar divisões entre atributos numéricos, de forma a simplificar as regras utilizadas (Mehta, Rissanen and Agrawal, 1995). Por sua vez, o fator de confiança diz respeito à confiança associada à utilização de *pruning*. Fatores de confiança mais baixos correspondem a um maior *pruning*, enquanto que valores superiores a 0.5 originam a desativação deste mecanismo (Gerdes, Galar and Scholz, 2016). Quanto ao número mínimo de objetos e ao número de *folds*, o primeiro relaciona-se com o número mínimo de instâncias por folha e o número de *folds* com a quantidade de dados utilizada para *pruning*. Dito de outro modo, o número mínimo de instâncias por folha indica o número mínimo de registos que é necessário assegurar-se para que uma dada regra seja considerada na árvore e, por sua vez, o número de *folds* tem a ver com o número de intervalos em que se dividem os dados, sendo um desses intervalos utilizado para *pruning* e os restantes para o crescimento da árvore (Sharma, Ghosh and Joshi, 2013).

Tabela 4 – Valores a otimizar nos parâmetros do algoritmo J48

Parâmetro	Valores a Variar
Fator de confiança	0.10, 0.25, 0.40
Nº mínimo de objetos	2, 4, 6, 10, 20
Reduced error pruning	V, F
Unpruned	V, F
Nº de *folds*	3, 6
Correção MDL	V, F
Total de simulações: 360	

Em relação ao algoritmo RF, os seus principais parâmetros são o número de árvores, a profundidade máxima de cada uma delas, o número de atributos e o *break ties*. Neste caso, o número de árvores refere-se ao número total de árvores gerado pelo algoritmo e, quanto maior for este número, melhor será, à partida, o seu desempenho. No entanto, para valores elevados, o ganho poderá ser insignificante quando comparado com o acréscimo incorrido no tempo de computação. No que diz respeito ao número de atributos, este valor é refente à quantidade máxima de atributos a integrar em cada uma das derivações das árvores criadas. Relativamente à profundidade máxima, este critério corresponde ao número máximo de níveis que as árvores podem conter, sendo o valor "0" representativo de árvores ilimitadas. Já o *break ties* é um parâmetro *booleano* que, caso se encontre ativo, elimina aleatoriamente atributos que exibam uma qualidade preditiva idêntica à de outros. Na **Tabela 5** mostram-se os valores que se fizeram variar para cada um destes campos, de forma a apurar-se a melhor configuração para esta técnica.

Tabela 5 - Valores a otimizar nos parâmetros do algoritmo RF

Parâmetro	Valores a Variar
Nº de árvores	20, 50, 80, 100
Profundidade máxima	0, 10, 25
Nº de atributos	0, 5, 10
Break ties	V, F
Total de simulações: 144	

Quanto ao NB, este algoritmo foi otimizado tendo em conta os parâmetros relativos ao *kernel* e à discretização supervisionada, e os valores que se fizeram variar estão representados na **Tabela 6**. Por regra, o NB assume que os dados apresentam uma distribuição normal, o que nem sempre sucede e, nestas situações, aos atributos numéricos pode ser aplicado um estimador da densidade de Kernel, de forma a melhorar o resultado preditivo. Outra opção que se pode selecionar, também para os atributos numéricos, é a discretização supervisionada, que tem como função convertê-los em atributos nominais. (John and Langley, 2013; Witten *et al.*, 2016) Evidentemente, não se pode utilizar a discretização supervisionada em simultâneo com estimadores da densidade de Kernel.

Tabela 6 - Valores a otimizar nos parâmetros do algoritmo NB

Parâmetro	**Valores a Variar**
Kernel	V, F
Discretização sup.	V, F
Total de simulações: 6	

A técnica KNN pode ser otimizada através de ajustes aos parâmetros *k*, *CrossValidate*, distância ponderada e ao método de pesquisa do vizinho mais próximo. O primeiro parâmetro, *k*, diz respeito ao número de vizinhos mais próximos a ser considerado e, em princípio, quanto maior for o seu valor, menor será a sensibilidade ao ruído. No entanto, este parâmetro requer uma otimização cuidada para que sejam apresentados bons resultados preditivos e, ao mesmo tempo, tem de existir um compromisso entre o parâmetro *k* escolhido e o tempo de execução do algoritmo, dado que altos valores de *k* se podem refletir em tempos computacionais excessivamente elevados. Em relação ao *CrossValidate*, o modo de funcionamento deste parâmetro consiste em inferir-se, de uma forma automática, o valor do número de vizinhos mais próximos. Para tal, a estimativa deste número é efetuada com recurso a mecanismos de *cross-validation*, fazendo-se variar o número de vizinhos mais próximos entre 1 e o valor de *k*. Assim, neste caso, sempre que a opção *CrossValidate* seja "verdade", o parâmetro *k* funciona, não como o valor exato do número de vizinhos mais próximos a considerar na execução do algoritmo, mas sim como o limite máximo a ser considerado na determinação automática do seu valor. Por sua vez, a distância ponderada é também um parâmetro intrínseco ao KNN, que se relaciona com a atribuição de pesos à distância compreendida entre a instância de teste e os seus vizinhos mais próximos. No *software* Weka, nas configurações deste algoritmo, além da opção selecionada por defeito, em que esta distância não é ponderada, dispõe-se de mais duas

fórmulas distintas para se associarem pesos às distâncias. O último parâmetro a ser configurado é o que se relaciona com a escolha do método de pesquisa do vizinho mais próximo e, para isso, o Weka possibilita a realização de testes com recurso a diferentes algoritmos. (Witten *et al.*, 2016) Os valores que se fizeram variar para a otimização dos parâmetros inerentes ao KNN são os que se indicam na **Tabela 7**.

Tabela 7 - Valores a otimizar nos parâmetros do algoritmo KNN

Parâmetro	**Valores a Variar**
k	10, 20, 50
CrossValidate	V, F
Distância ponderada	Não, 1/Dist, 1-Dist
Algoritmo de pesquisa do vizinho mais próximo	*LinearNNSearch*, *BallTree*, *FilteredNSearch*, *KDTree*
Total de simulações: 72	

No que toca ao MLP, este algoritmo permite que sejam otimizados os valores dos parâmetros referentes às camadas escondidas, ao tempo de treino, ao filtro que transforma valores nominais em binários, ao *decay*, ao *learning rate* (LR) e ao *momentum*. Os valores que se fizeram variar, para cada um destes parâmetros, estão indicados na **Tabela 8**. Como nota, realça-se o facto de se ter considerado o filtro *Normalize Attributes* sempre ativo em todas as simulações, devido à heterogeneidade verificada entre os valores dos atributos. O campo relativo às camadas escondidas representa o número de camadas escondidas e o número de neurónios existente em cada uma delas. O seu valor pode ser inserido na forma de um formato numérico ou, em alternativa, pode introduzir-se uma letra com um significado pré-definido no Weka. No caso das simulações efetuadas para o *tuning* deste algoritmo, tal como se observa da **Tabela 8**, considerou-se apenas uma camada escondida, fazendo-se variar o valor do número de neurónios em "*a*" e "*o*". Neste caso, o valor "*a*" é dado por $a = (n^o\ atributos + n^o\ classes)/2$ e, por isso, assume o valor "7" no cenário I e o valor "6" no cenário II, enquanto "o", que corresponde ao número total de classes, é "2" em ambos os cenários. O Anexo I contém um esquema do esqueleto das redes neuronais simuladas para o cenário I. No caso do cenário II, as redes testadas apresentam uma arquitetura idêntica, diferindo apenas nos atributos da camada de *input* e no número de neurónios da camada escondida quando o seu valor é configurado para

"*a*". Relativamente ao tempo de treino, este valor é referente ao número de *epochs*, ou seja, ao número de ciclos que o algoritmo irá efetuar, percorrendo, em cada um deles, todos os dados de treino. Quanto ao *Nominal para Binário*, este filtro converte, quando ativo, os valores nominais em valores binários. Os parâmetros LR e *momentum* representam valores associados aos pesos do algoritmo. Por sua vez, o parâmetro *decay* está relacionado com o LR, na medida em que é responsável pela diminuição do seu valor. Assim, quando o *decay* assume um valor verdadeiro, este parâmetro compele a que o valor do LR seja continuamente atualizado e dado pela divisão do seu valor inicial com o número da *epoch* a ser executada no momento. Como valores altos de LR necessitam de menos *epochs* para atingirem soluções, normalmente abaixo do "ideal", e como valores mais baixos requerem mais *epochs* para que os resultados se aproximem mais dos reais, o parâmetro *decay* funciona como um ajuste à técnica, equilibrando os valores de LR consoante a fase em que se encontra a execução do algoritmo. Deste modo, a ativação do *decay* promove a convergência do algoritmo para uma solução "ideal" e, com isso, a sua utilização pode proporcionar melhores desempenhos (Witten *et al.*, 2016).

Tabela 8 - Valores a otimizar nos parâmetros do algoritmo MLP

Parâmetro	**Valores a Variar**
Camadas escondidas	a, o
Tempo de treino	100, 300
Nominal para Binário	V, F
Decay	V, F
LR	0.1, 0.2, 0.3
Momentum	0.05, 0.10, 0.20
Total de simulações: 288	

Uma vez definidos os parâmetros e os valores a otimizar em cada uma das técnicas candidatas ao cálculo dos índices, foi executado um total de 870 simulações tendo-se em conta os dois cenários criados, cujos resultados se sintetizam nos Anexos II e III. Os resultados mostraram que as técnicas se comportavam de forma idêntica quando sujeitas aos mesmos ajustes dos parâmetros em ambos os cenários.

No que diz respeito ao algoritmo J48, as principais observações retiradas dos resultados obtidos, para ambos os cenários, são as que se listam:

- À medida que o número mínimo de objetos aumenta, a acurácia também aumenta de forma ligeira, salvo em algumas exceções verificadas no cenário II.
- Uma árvore submetida a *pruning* (*unpruned* = F) apresenta melhores resultados preditivos do que outra árvore não sujeita a esta prática, com os restantes parâmetros iguais. Este ponto mostra que, sem *pruning*, o algoritmo cria árvores com complexidades excessivas, o que prejudica a previsão.
- Nota-se que a diferença entre as acurácias de árvores sujeitas aos mesmos parâmetros à exceção do *pruning* diminui, significativamente, à medida que o número mínimo de objetos aumenta, de forma igual, em ambas as árvores. A explicação para isto ocorrer pode dever-se ao facto de o aumento do número mínimo de objetos estar a substituir o papel do *pruning*, ao fazer diminuir as ramificações da árvore não submetida a esta prática. Assim, neste caso, o aumento do número mínimo de objetos pode estar a diminuir o grau de complexidade da árvore não sujeita a *pruning* e a aproximá-lo ao da árvore sujeita a esta prática.
- Como seria de esperar, em árvores não sujeitas a *pruning*, os resultados são independentes do fator de confiança e do número de *folds*, uma vez que estes parâmetros estão relacionados com o *pruning*.
- Em árvores submetidas a *pruning*, um aumento do fator de confiança traduz-se numa ligeira diminuição do valor da acurácia.
- O número de *folds* praticamente não tem interferência nos valores da acurácia, nos falsos negativos e no tempo de execução.
- O MDL ativo melhora, consideravelmente, os resultados.
- A taxa de falsos negativos não apresenta um padrão de comportamento típico e os seus valores são, por norma, relativamente homogéneos. No entanto, existe uma ligeira melhoria desta taxa quando os parâmetros *reduced error pruning* e MDL estão ativos.
- Os tempos de execução do algoritmo são maiores quando o número mínimo de objetos é baixo ou quando o parâmetro MDL não está ativo.
- De uma forma geral, é preferível que o *reduced error pruning* esteja ativo quando o fator de confiança é superior a 0.25 ou quando o número mínimo de objetos é igual ou superior a 10. Nas restantes situações, os melhores resultados ocorrem quando este parâmetro está desativado.

Deste modo, concluiu-se que os parâmetros *unpruned* e MDL deveriam assumir os valores de "falso" e "verdadeiro", respetivamente, independentemente dos restantes parâmetros, e que o fator de confiança deveria ser, preferencialmente, 0.10. Além disso, verificou-se ainda que o aumento do número mínimo de objetos originava um aumento no valor da acurácia e uma diminuição no tempo de computação, tendo-se, por isso, optado pelo valor "20" para este número. Como se trata de um número mínimo de objetos elevado, atribuiu-se o valor "verdade" ao *reduced error pruning*. Por sua vez, o número de *folds* foi ajustado para 6, no caso do cenário I, e para 3 para o cenário II, uma vez que, tendo em conta os parâmetros já selecionados, estes correspondiam aos valores que se refletiam em ligeiros ganhos de acurácia e de tempos de execução.

Para o algoritmo RF, os resultados obtidos, considerando os dois cenários criados, permitiram apurar que:

- Quando o número de árvores é incrementado, o valor da acurácia aumenta, embora a partir de 80 árvores se verifique uma estagnação no seu valor. Desta forma, quando o número de árvores é superior a 80, é atingida a saturação e, por conseguinte, mais árvores geradas não resultam em melhorias nos resultados preditivos. Por outro lado, o aumento no número de árvores implica maiores tempos de execução e aumenta também, de forma ligeira, a taxa de falsos negativos.
- Nos casos em que a profundidade máxima das árvores aumenta, a acurácia e a taxa de falsos negativos diminuem e, por sua vez, o tempo de computação aumenta. Este aumento de tempo faz-se sentir, sobretudo, nas situações em que o número de árvores é elevado.
- Um maior número de atributos a considerar provoca, por norma, uma quebra ligeira nos valores de acurácia e não tem correspondência direta com o número de falsos negativos. Além disso, perante situações em que o número de árvores e a profundidade de cada uma delas sejam elevados, esta subida pode originar um maior tempo de execução.
- O parâmetro *break ties* não aparenta influenciar os resultados.

Do exposto, inferiu-se que o número de árvores ótimo e o número de atributos a serem considerados deveria ser 80 e 0, respetivamente. Para além disso, notou-se que a profundidade máxima de cada uma das árvores não deveria corresponder a um valor extremo, dado o *trade-off* existente entre a acurácia, os falsos negativos e o tempo de execução. Deste modo, a profundidade máxima de cada árvore foi definida como "25". Para a escolha do valor lógico do

break ties, tiveram-se em conta os parâmetros definidos, e verificou-se que, no cenário II, caso o seu valor fosse "verdade", os valores registados quer para a acurácia, quer para os falsos negativos, eram ligeiramente melhores. No caso do cenário I, a sua influência é também praticamente insignificante, sendo que o valor "verdade", neste caso, aumentou ligeiramente a acurácia e a taxa de falsos negativos. Assim, considerou-se, em ambos os cenários, o valor "verdade" para o *break ties*.

O algoritmo NB, devido à sua simplicidade, apenas permite a otimização de dois parâmetros. Assim, analisando-se os resultados, deduz-se que:

- A melhor situação é recorrer à discretização supervisionada, que proporciona os melhores resultados em termos de acurácia e de falsos negativos, embora estes resultados sejam semelhantes ao que se obtêm quando se utilizam estimadores de Kernel.
- O pior caso é observado quando não se utiliza discretização supervisionada nem estimadores de Kernel.
- O tempo de execução, independentemente dos restantes parâmetros, é praticamente nulo em todas as simulações.

Deste modo, para este algoritmo, a otimização dos parâmetros que conduziam aos melhores resultados não apresentou um carácter subjetivo e, por isso, a melhor escolha recaiu, em ambos os cenários, por se utilizar discretização supervisionada.

Quanto ao KNN, as principais considerações retiradas são também comuns a ambos os cenários:

- O valor de *k*, à medida que aumenta, origina, por regra, maiores valores de acurácia, de falsos negativos e de tempo de execução. No entanto, em relação à taxa de falsos negativos, observa-se uma exceção no cenário II, quando o *CrossValidate* está ativo e, por outro lado, verifica-se também que, para valores de *k* superiores a 20, o aumento de acurácia e de tempo de computação é pouco significativo.
- O *CrossValidate* não tem qualquer influência ao nível da percentagem de acerto, embora, em relação aos falsos negativos, as melhores taxas sejam, normalmente, detetadas quando este parâmetro assume o valor "falso", salvo raras exceções. Além disso, neste caso, o tempo de execução observado também é substancialmente inferior em relação ao que acontece quando o *CrossValidate* está ativo.

- O valor da acurácia praticamente não é afetado pela ponderação da distância, mas a taxa de falsos negativos e o tempo computacional são inferiores nas situações em que não é utilizada qualquer ponderação.
- Os algoritmos de pesquisa do vizinho mais próximo em análise originam, praticamente, os mesmos valores de acurácia e de falsos negativos, excetuando-se o algoritmo *FilteredNNSearch* que, normalmente, está associado a menores valores de acurácia e maiores taxas de falsos negativos. Importa ainda notar que, por vezes, o *KDTree* apresenta resultados ligeiramente melhores do que o dos restantes algoritmos ao nível da acurácia e dos falsos negativos.
- O algoritmo *KDTree* é o que conduz, visivelmente, aos menores tempos de execução e, em contrapartida, o *LinearNNSearch* é o responsável pelos maiores tempos.

Desta forma, definiu-se que, para o KNN, o valor de *k* mais adequado a aplicar era 20, de modo a maximizar-se a acurácia do modelo e, ao mesmo tempo, garantir-se valores aceitáveis de taxas de falsos negativos. Além disso, optando-se pela desativação do filtro *CrossValidate* e escolhendo-se o algoritmo *KDTree*, este valor para *k*, apesar de elevado, quando conjugado com estes parâmetros, não se reflete em tempos computacionais elevados. Adicionalmente, considerou-se que a distância não deveria ser ponderada.

Em relação ao algoritmo MLP, para os cenários I e II, verificou-se que:

- Quando o número de neurónios da camada escondida é superior, a acurácia apresenta melhores resultados, embora o tempo de execução seja mais elevado. Por outro lado, os falsos negativos não aparentam relacionar-se com este parâmetro.
- A subida do tempo de treino de 100 para 300 aumenta, em cerca de três vezes, o tempo de computação. No entanto, este tempo de treino não aparenta ter correlação com a percentagem de acerto nem com o número de falsos negativos.
- A utilização do filtro que transforma os valores nominais em binários praticamente não influencia os resultados, embora, de uma forma geral, melhore ligeiramente os valores relativos à acurácia e aos falsos negativos e provoque, superficialmente, um agravamento no tempo de computação.
- Normalmente, os melhores valores de acurácia e de falsos negativos são conseguidos pela desativação do *decay*. Contundo, para valores de LR superiores a 0.2, a acurácia melhora com a ativação deste parâmetro, apesar de a taxa de falsos negativos ser igualmente pior nesta situação.

- Geralmente, quando o *decay* está ativo, aumentos no LR ou no *momentum* traduzem-se em melhores resultados preditivos. No entanto, quando desativo, verifica-se o oposto e maiores valores de LR ou do *momentum* conduzem a piores resultados ao nível da acurácia e da taxa de falsos negativos.

Em suma, para o algoritmo MLP, consideraram-se "*a*" neurónios para a camada escondida, com o filtro que converte os nominais em binários ativo, para um tempo de treino 100. Assim, a conjugação destes parâmetros tornou viável o tempo de computação que, independentemente dos restantes, seria sempre da ordem dos segundos. Quanto aos restantes parâmetros, o *decay* foi desativado, de modo a que não se originassem taxas de falsos negativos mais elevadas e, por sua vez, ao LR e ao *momentum* atribuíram-se os valores mais baixos testados.

Do cômputo da análise efetuada, os valores atribuídos aos parâmetros examinados, para cada uma das técnicas em estudo, são os que se sumariam na **Tabela 9**.

Tabela 9 - Sumário da configuração dos parâmetros para as técnicas em estudo

J48	RF	NB	KNN	MLP
F. Confiança: 0.10	Nº Árvores: 80	Kernel: F	*k*: 20	Camadas Escondidas: *a*
Nº Mín. Obj.: 20	Profundidade Máxima: 25	Discretização Supervisionada: V	*CrossValidate*: F	Tempo Treino: 100
Reduced Error Pruning: V	Nº Atributos: 0		Distância Ponderada: Não	Nominal para Binário: V
Unpruned: F	*Break Ties*: V		Algoritmo de Pesquisa: *KDTree*	*Decay*: F
Nº *Folds*:				LR: 0.1
Cenário I: 6				*Momentum*: 0.05
Cenário II: 3				
Correção MDL: V				

Deste modo, tendo-se em conta a implementação do Weka ilustrada na **Figura 12**, a etapa seguinte consistiu na substituição dos valores por defeito dos parâmetros de cada uma das técnicas para os indicados na **Tabela 9**. Assim, executando-se os ficheiros criados para os dois cenários, foi possível gerar os modelos e interpretar os seus resultados.

A nível visual, os resultados relativos ao J48 foram os mais fáceis de analisar, uma vez que puderam ser examinados graficamente, na forma de uma única árvore, na qual os atributos estavam hierarquicamente representados. Assim, nesta técnica, observou-se que, quer no cenário I, quer no II, a pressão arterial era o atributo do topo da árvore e, por conseguinte, o mais relevante do conjunto em análise. Além disso, também se verificou que, em ambos os

cenários, o colesterol total e a glicose rápida eram fatores com uma alta relevância preditiva. As árvores criadas apresentaram uma grande complexidade, com um tamanho 289 no cenário I e 173 no cenário II, e eram formadas por um grande conjunto de regras. No caso do primeiro cenário, a árvore gerada continha 158 regras (folhas), ao passo que a do cenário II se regia por um total de 96. A título de exemplo, na **Figura 13**, apresenta-se um excerto, da parte superior da árvore gerada pelo algoritmo J48, no cenário II.

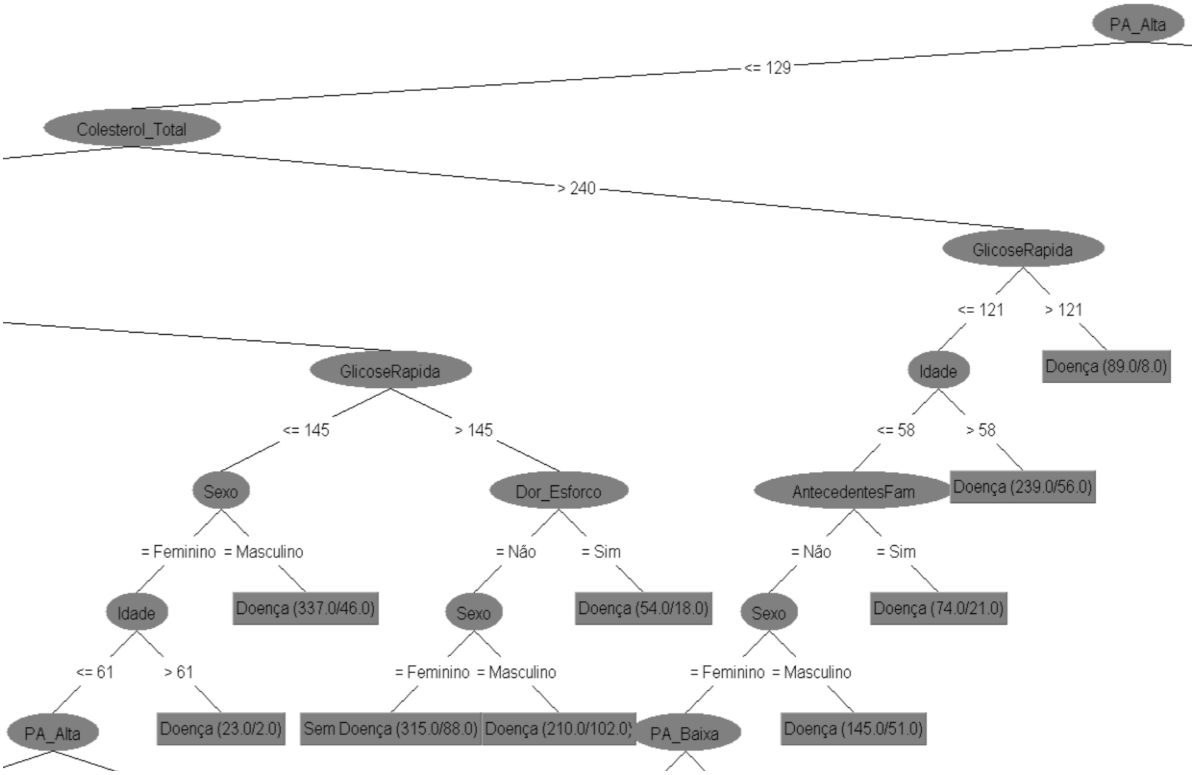


Figura 13 - Excerto da árvore gerada pelo algoritmo J48 (cenário II).

Importa notar, através da visualização da figura, que uma folha, representada por um retângulo, corresponde ao nodo final de uma possível trajetória na árvore de decisão. Desta forma, as folhas da árvore representam regras e correlações entre atributos. Como exemplo, observa-se, através da folha localizada mais à direita, que uma das regras a que o algoritmo J48 (cenário II) recorre para classificar um registo como sendo do tipo "Doença" é: "***SE*** *pressão arterial alta <= 129* ***E SE*** *colesterol total >240* ***E SE*** *glicose rápida > 121* ***ENTÃO*** *doença*". Num panorama bastante geral, dada a complexidade das árvores, as regras pareceram plausíveis tendo em conta os resultados que seriam expectáveis da literatura.

Contrariamente ao J48, os modelos criados pelas restantes técnicas foram de mais difícil interpretação visual. No caso do RF, como neste algoritmo se geraram bastantes árvores, cada uma delas com uma complexidade ainda superior à do J48 por não terem sido sujeitas a *pruning*, e o resultado preditivo depende do conjunto total de árvores, as conclusões acerca dos

parâmetros que se consideraram ser os que mais condicionam a previsão tornaram-se de difícil entendimento. Por outro lado, para o NB e para o MLP não foi possível observar os resultados finais da criação dos modelos de uma forma gráfica. No primeiro caso, somente puderam ser visualizadas as frequências consideradas para cada *label* de cada atributo, tendo-se em conta as classes "Doença" e "Sem Doença", ao passo que, no MLP, os resultados refletiram apenas os pesos, por nodo, determinados para a conceção do modelo. Desta forma, a partir destas técnicas não foi possível efetuar uma interpretação visual dos modelos criados, uma vez que se retornaram apenas resultados intermédios, que serviram de base para a sua criação. Além disso, dado o modo de funcionamento da técnica KNN, não foram gerados resultados gráficos para este algoritmo, tendo-se apenas comprovado que o modelo tinha sido construído para 20 vizinhos mais próximos. No Anexo IV, estão representados excertos dos resultados que se obtiveram para as técnicas RF, NB e MLP, considerando-se o conjunto de atributos do cenário I.

Com o intuito de se proceder a uma análise detalhada acerca da viabilidade de cada uma destas técnicas no processo de deteção de DCV, foi construída a matriz de confusão detalhada na **Tabela 10**.

Tabela 10 - Matriz de confusão relativa a cada uma das técnicas em estudo para a previsão de DCV

			Previsão									
			Doente					**Saudável**				
			J48	RF	NB	KNN	MLP	J48	RF	NB	KNN	MLP
Realidade	**Cenário I**	**Doente**	17340	17561	16159	17123	17496	7646	7425	8827	7863	7490
		Saudável	5552	5960	5041	6593	5650	22553	22145	23064	21512	22455
	Cenário II	**Doente**	17198	17296	15936	17518	17084	7788	7690	9050	7468	7902
		Saudável	5462	6612	5028	6610	5583	22643	21493	23077	21495	22522

Num sentido mais lato, a matriz obtida permitiu constatar que o modelo resultante da aplicação do algoritmo RF, construído com base nos atributos do cenário I, era aquele que identificava um maior número de doentes corretamente, enquanto o modelo associado ao NB, relativo ao cenário II, era o que detetava a maior parte dos indivíduos saudáveis de forma acertada. Por outro lado, os algoritmos que mostraram ter um pior desempenho no reconhecimento de indivíduos doentes e saudáveis foram o NB e o RF, respetivamente, ambos relativos ao cenário II.

No contexto em estudo, o modelo preditivo a ser selecionado deve, não só apresentar valores corretos, como também identificar a maioria das pessoas doentes acertadamente, ainda que, por vezes, isso implique uma classificação errada de pessoas saudáveis. Assim, as métricas que se consideraram determinantes para a seleção do melhor algoritmo foram a acurácia e a sensibilidade. A sensibilidade é uma métrica que apura a eficácia na determinação de indivíduos com doença, no universo de indivíduos doentes e, por conseguinte, está inversamente relacionada com o número de pessoas doentes que são incorretamente classificadas. Relembre-se que, para o processo de *tuning* das técnicas, os aspetos que se consideraram para se apurarem os melhores parâmetros tinham sido a acurácia e a taxa de falsos negativos. No entanto, neste caso, para o processo de avaliação das técnicas, a taxa de falsos negativos não é a métrica mais indicada, uma vez que, ao contrário do procedimento anterior, cujos resultados serviram apenas de apoio para efeitos comparativos, este processo requer uma ponderação dos valores absolutos da acurácia e dos falsos negativos. Deste modo, caso o universo de indivíduos saudáveis seja significativamente maior do que o de não saudáveis, a taxa de falsos negativos pode resultar num valor muito baixo e, na realidade, o número de doentes incorretamente classificados ser elevado face ao número total de doentes. Assim, a melhor forma de se ter em consideração esta situação é utilizar a sensibilidade em detrimento da taxa de falsos negativos.

Para se avaliarem os resultados indicados na matriz de confusão da Tabela 10 em termos de acurácia e de sensibilidade, é preciso utilizar a **Equação 4** e a **Equação 5**, por essa ordem. Nestas equações, VP representa o número de verdadeiros positivos, VN diz respeito ao número de verdadeiros negativos, FP é referente aos falsos positivos e, por sua vez, a sigla FN está associada aos falsos negativos. No entanto, além da matriz de confusão, os ficheiros de texto de cada modelo, resultantes do *TextSaver* do Weka, também retornam os valores de acurácia e de sensibilidade já determinados.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4}$$

$$Sensibilidade = \frac{VP}{VP + FN} \tag{5}$$

De forma a estabelecer-se um mecanismo automático para a seleção do melhor modelo, de acordo com as métricas de avaliação definidas, recorreu-se ao Spoon. No caso do presente trabalho, considerou-se que os modelos que melhor respondem às necessidades do DM são aqueles que somam as maiores pontuações, tendo em conta os valores de acurácia e de

sensibilidade inerentes a cada um. Desta forma, considerou-se que o *score* de cada modelo é dado pela média dos seus valores de acurácia e de sensibilidade, assumindo-se, por isso que, na previsão de DCV, estas duas métricas têm igual relevância. Dito de outro modo, julgou-se igualmente importante a capacidade de acerto dos modelos e, ao mesmo tempo, a sua capacidade de cobertura na identificação dos indivíduos com doença.

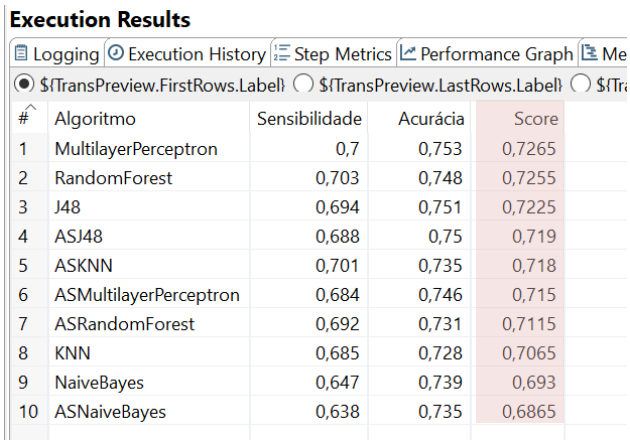
Assim, a partir dos ficheiros resultantes do *TextSaver* do Weka, cada um deles identificado pelo nome da técnica e pelo cenário a que diz respeito, implementou-se o sistema ilustrado na **Figura 14**.

Figura 14 – Implementação, no Spoon, para a seleção do melhor modelo para o caso em estudo.

Numa primeira fase, o sistema criado reúne, isoladamente, os ficheiros que contêm os resultados obtidos para cada um dos modelos e, para cada um deles, filtra apenas o conteúdo referente aos valores de sensibilidade e de acurácia. O passo seguinte consistiu no acréscimo de uma nova coluna com a identificação da técnica e do cenário a que os resultados pertenciam, para poderem ser posteriormente compilados e comparados. Para isso, foi preciso determinar-se e ordenar-se, de forma decrescente, o *score* de cada modelo que, tal como explicitado, corresponde à média entre os valores de sensibilidade e de acurácia. Assim, considerou-se que o modelo mais indicado para a previsão de DCV seria o dotado de maior *score*.

Aplicando-se o sistema desenvolvido ao caso de estudo, os resultados mostraram que o modelo gerado pelo MLP, criado com base nos atributos do cenário I, correspondia à técnica com maior score (0.7265) e, por conseguinte, à mais apropriada a ser utilizada. Os resultados

que se obtiveram são os representados na **Figura 15** e permitiram estabelecer um *ranking* ordenado de acordo com os melhores modelos a usar em previsões de DCV: MLP ≻ RF ≻ J48 ≻ ASJ48 ≻ ASKNN ≻ ASMLP ≻ KNN ≻ NB ≻ ASNB.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Me

${TransPreview.FirstRows.Label} ${TransPreview.LastRows.Label} ${Tra

#	Algoritmo	Sensibilidade	Acurácia	Score
1	MultilayerPerceptron	0,7	0,753	0,7265
2	RandomForest	0,703	0,748	0,7255
3	J48	0,694	0,751	0,7225
4	ASJ48	0,688	0,75	0,719
5	ASKNN	0,701	0,735	0,718
6	ASMultilayerPerceptron	0,684	0,746	0,715
7	ASRandomForest	0,692	0,731	0,7115
8	KNN	0,685	0,728	0,7065
9	NaiveBayes	0,647	0,739	0,693
10	ASNaiveBayes	0,638	0,735	0,6865

Figura 15 - *Scores* relativos a cada um dos modelos analisados.

Para a validação dos resultados obtidos, foram utilizados os dados de teste, do *dataset* que continha 5.000 registos, de forma a avaliar-se se, perante novos dados, os valores de sensibilidade e de acurácia permaneciam idênticos. Assim, depois de devidamente tratados estes registos, determinaram-se os valores de sensibilidade e de acurácia, para os novos dados, utilizando-se os mesmos modelos. Note-se que, neste caso, o procedimento de tratamento de dados efetuado foi ligeiramente diferente do anterior, na medida em que não foram considerados os filtros que poderiam descartar registos. Na **Tabela 11** estão expostos os resultados obtidos.

Tabela 11 – Quadro comparativo entre as fases de treino e validação e a de teste dos modelos

Algoritmo	Sensibilidade	Δ Sensibilidade	Acurácia	Δ Acurácia
MLP	0.7117	↑ 1.17 %	0.7232	↓ 2.98 %
RF	0.7093	↑ 0.63 %	0.7288	↓ 1.92 %
J48	0.6895	↓ 0.45 %	0.7192	↓ 3.18 %
J48 (AS)	0.7130	↑ 2.50 %	0.7294	↓ 2.06 %
KNN (AS)	0.7121	↑ 1.11 %	0.7074	↓ 2.76 %
MLP (AS)	0.6980	↑ 1.40 %	0.7372	↓ 0.88 %
RF (AS)	0.7024	↑ 1.04 %	0.7206	↓ 1.04 %
KNN	0.6931	↑ 0.81 %	0.7032	↓ 2.48 %
NB	0.6595	↑ 1.25 %	0.7418	↑ 0.28 %
NB (AS)	0.6445	↑ 0.65 %	0.7380	↑ 0.30 %

Dos registos tabelados, observou-se que os valores de sensibilidade e de acurácia obtidos para os dados de teste apresentaram sempre variações inferiores a 5% e, por isso, comprovou-se a adequabilidade dos modelos construídos a outros *datasets* distintos, uma vez que não revelaram a existência de *overfitting*. Desta forma, os *scores* obtidos na **Figura 15** foram validados e, neste caso, o modelo selecionado foi o relativo à técnica MLP (cenário I).

Terminada esta fase, os objetivos do DM foram cumpridos e a etapa seguinte passou por se efetuar uma avaliação do modelo face aos objetivos de negócio definidos.

3.5 Avaliação

Para que se pudessem satisfazer os objetivos do negócio, foi preciso criar-se um mecanismo que permitisse estimar os valores dos índices de bem-estar cardíaco. Relativamente aos índices, convencionou-se que a sua escala seria expressa entre -5 e 5 e que cada valor corresponderia a uma cor. Neste caso, consideraram-se quatro cores possíveis, com o seguinte significado:

- **Verde** (valores entre 2.5 e 5) – o risco de desenvolvimento de DCV é reduzido. Os indivíduos devem manter os seus hábitos, de forma a que os fatores de risco não sejam aumentados.
- **Amarelo** (valores entre 0 e 5) – o risco de desenvolvimento de DCV é moderado e, por isso, é necessário algum controlo regular e agir de acordo com comportamentos que promovam a redução dos fatores de risco.
- **Laranja** (valores entre -2.5 e 0) – o risco de desenvolvimento de DCV é elevado. Aconselha-se a que, nesta situação, os indivíduos recorram a especialistas em DCV, de forma a realizarem exames de diagnóstico mais específicos.
- **Vermelho** (valores entre -5 e -2.5) – o risco de desenvolvimento de DCV é muito elevado. Deve-se procurar com urgência especialistas em DCV e realizar exames de diagnóstico específicos.

Os valores dos índices podem ser apurados pela implementação de um sistema que converta, para cada registo, os valores probabilísticos associados às classes "Doença" e "Sem Doença" em índices, recorrendo, para isso, ao modelo selecionado na fase da *Modelação*. No entanto, é preciso notar que esta conversão não deve ser efetuada de uma forma linear, uma vez que se pretende que os índices funcionem como meios de alerta, capazes de detetarem

antecipadamente o surgimento de DCV. Assim, o índice associado a um determinado registo deverá diminuir de forma exponencial perante um aumento no risco de doença.

Para se implantar o sistema descrito, considerou-se que os riscos de desenvolvimento de DCV de 0 e de 100% estão associados a valores de índice 5 e -5, respetivamente. Além disso, após várias simulações, definiu-se que o limiar entre as cores laranja e vermelho deveria corresponder a valores de risco de 55%. Deste modo, tendo-se três pontos selecionados, foi possível delinear uma curva exponencial que os intersetasse e cuja equação permitisse inferir quaisquer outros valores de índices. A curva construída, dada pela **Equação 6**, é a que se representa na **Figura 16**.

$$f(x) = -7.052755 + 12.05276 \times e^{-1.77011 \times (x)} \quad (6)$$

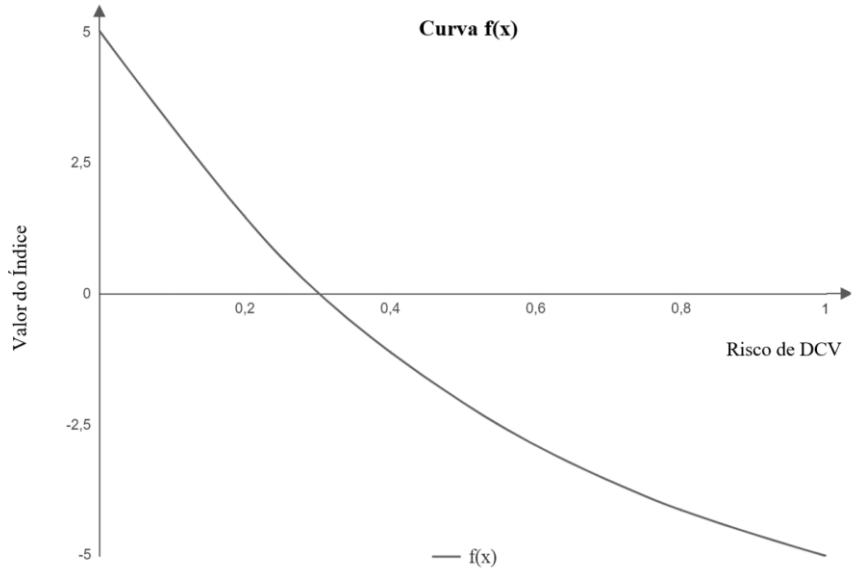


Figura 16 - Curva de correspondência entre o grau de risco de DCV e o valor do índice.

De forma a avaliar-se se o modelo cumpria os objetivos iniciais, a partir dos dados de teste da etapa anterior, para cada um dos registos estimou-se o seu valor de índice e a sua correspondente cor. Assim, e dado que o *dataset* continha dados classificados, foi elaborado o gráfico da **Figura 17**, que permitiu identificar a percentagem de indivíduos doentes e não doentes previstos em cada uma das categorias, face ao total de registos em cada classe.

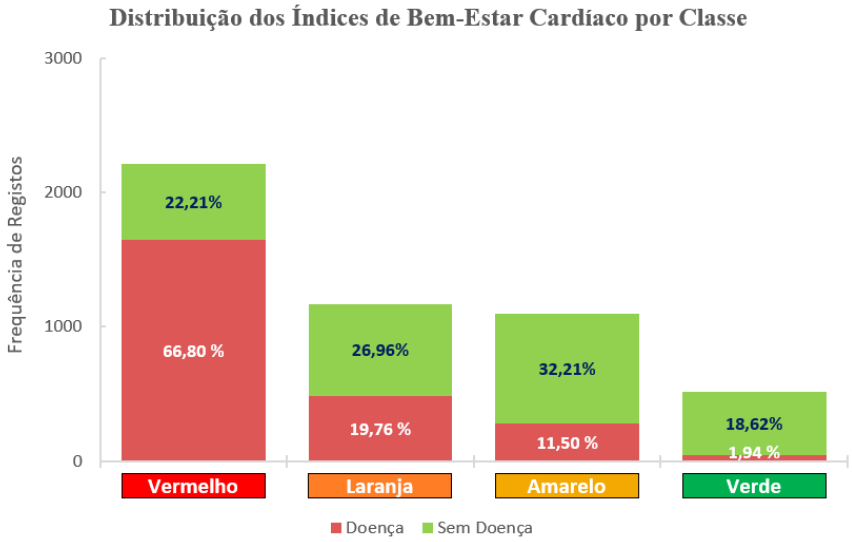


Figura 17 - Previsões da cor do índice associado a cada registo, de acordo com a classe.

Globalmente, verificou-se que a maioria dos indivíduos doentes se enquadrava na cor vermelha do índice e que, ao todo, 86.56% dos indivíduos doentes tinham associados índices de cor vermelha ou laranja. Pelo contrário, observou-se que 1.94% destes indivíduos tinham um índice de cor verde associado. O "ideal", neste caso, era que todos os indivíduos com índices de cor verde não tivessem DCV. No entanto, a percentagem de indivíduos doentes associada a esta cor verificou-se ser muito reduzida e inferior a 2%. A explicação mais plausível para se ter observado registos de indivíduos doentes com um índice verde reside no facto de alguma da informação do *dataset* ser de carácter subjetivo e poder estar errada, como é o caso do exercício físico, dos antecedentes familiares e do número de cigarros fumados. Além disso, os valores relativos à pressão arterial alta e baixa e aos dados das análises médicas não têm em conta os dados históricos e, consequentemente, podem refletir valores muito discrepantes em comparação aos do passado. É preciso ter ainda em conta que nem sempre a existência de baixos fatores de risco impede o surgimento deste tipo de doenças. Desta forma, considerou-se aceitável esta taxa de erro de cerca de 2%, uma vez que a cor verde não simboliza a existência de riscos nulos de desenvolvimento de DCV, mas sim de probabilidades muito reduzidas.

Quanto aos indivíduos sem doença, verificou-se que, ao todo, apenas 50.83% se inseria nas cores verde e amarelo. Deste modo, praticamente metade dos indivíduos saudáveis ficaram associados a índices laranja e vermelho. Este resultado pode ser considerado válido na medida em que diz respeito a pessoas com fatores de risco elevados de desenvolvimento de DCV, mas que não desenvolveram estas doenças e, deste modo, serve para as alertar e incentivar a consultarem profissionais de saúde especializados, para efetuarem exames de diagnóstico, de

forma a prevenir o seu surgimento. Além disso, os profissionais de saúde podem ainda informá-las acerca das melhores medidas a tomar para que os seus fatores de risco sejam diminuídos.

Outro modo complementar de se verificar a adequabilidade do índice proposto foi através da sua comparação com valores obtidos em simuladores *online*. Para isso, foram definidos os quatro tipos de perfis representados na **Figura 18** e, para cada um destes perfis, foi calculado o seu índice através do modelo MLP selecionado, que corresponde ao valor indicado na figura. Posteriormente, para os mesmos perfis, utilizando-se os simuladores *online*, obteve-se os resultados expostos na **Tabela 12**, que representam o risco de se vir a sofrer de algum tipo de evento cardiovascular grave nos próximos dez anos.

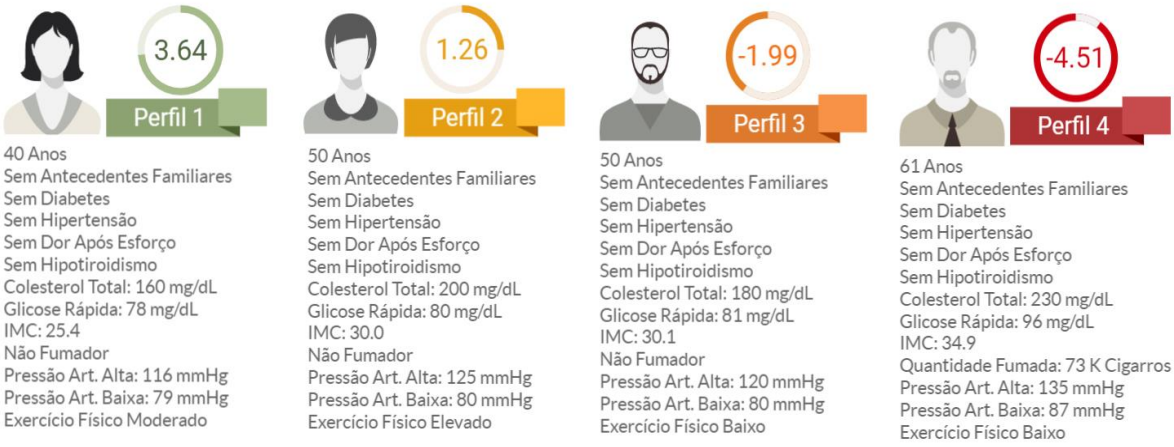


Figura 18 - Perfis-tipo para comparação dos índices calculados com os dos simuladores *online*.

Como estes simuladores não utilizavam os mesmos parâmetros preditivos e, em alguns deles, era requerida informação adicional que não estava indicada nos perfis selecionados, sempre que necessário, foram considerados os seguintes pressupostos:

- HDL de 40 mg/dL para os homens e de 50 mg/dL para as mulheres.
- Proteína C reativa com o valor de 1 mg/L.
- Etnia caucasiana.
- Sem utilização de tratamentos com estatina.
- Sem utilização de tratamentos à base de aspirina.

Os valores retornados pelos simuladores estavam expressos numa forma percentual e associados a níveis que se podiam considerar baixos quando o valor era inferior a 5%, ligeiros para valores entre 5 e 7.5%, médios quando o intervalo de valores era entre 7.5 e 20% e elevados para valores superiores a 20%.

Tabela 12 – Percentagem de risco de desenvolvimento de DCV a 10 anos pelos simuladores *online*

Simulador	Perfil 1	Perfil 2	Perfil 3	Perfil 4
Framingham (Lípidos)	1.7%	4.5%	7.9%	21.1%
Framingham (IMC)	2%	4.9%	9.8%	30.2%
ASCVD	0.4%	1.3%	3.4%	18.8%
Cuore	1%	0.9%	2.5%	13.5%
Reynolds	0.3%	1%	3%	15%

Da análise dos valores obtidos e da comparação destes resultados com o dos índices associados a cada perfil, constatou-se que, no caso do perfil 1, todos os simuladores retornavam um nível de risco muito baixo e, por isso, os resultados estavam em concordância com o valor do índice de 3.6 (zona verde). Tal como no perfil 1, em relação ao perfil 2, os valores obtidos também se situavam na gama de risco baixo, embora os de Framingham apresentassem valores próximos do nível de risco superior. Desta forma, tendo em consideração este aumento no grau de risco, pôde-se considerar adequado o índice positivo de 1.26 (zona amarela) obtido pelo modelo de DM. Posteriormente, analisando-se os resultados dos simuladores para o perfil 3, verificou-se que, apesar de estas ferramentas retornarem valores mais elevados em comparação com os primeiros dois perfis, a sua classificação continua a ser do tipo baixo risco. Apesar disso, os simuladores de Framingham já o associam a um perfil de risco médio de desenvolvimento de DCV, dois níveis acima dos anteriores simuladores. Por sua vez, o índice calculado com recurso ao MLP é negativo e praticamente igual a -2 (zona laranja). De uma forma análoga à justificação do índice do perfil 2, esta classificação foi considerada apropriada, uima vez que se trata de uma forma de incentivo para que os utilizadores melhorem o valor e a cor correspondente ao seu índice de bem-estar. Por último, em relação ao grau de risco do perfil 4, os simuladores de Framingham consideram-no elevado, enquanto os restantes o categorizam como sendo um risco médio. Do ponto de vista do índice proposto, o valor obtido foi de -4.51 e, por isso, os indivíduos enquadrados neste perfil foram atribuídos à zona de mais elevado risco (zona vermelha). Mais uma vez, por estar em consonância com os resultados de maior risco, registados pelos simuladores de Framingham, considerou-se que o seu valor correspondia ao pretendido.

Assim, constatou-se que os índices calculados com suporte ao modelo MLP selecionado, apresentavam correspondências com os simuladores *online*, sobretudo os de Framingham. Por outro lado, quando a correspondência não era total, o índice apresentou valores do lado da segurança, promovendo, deste modo, a prevenção de surgimento de DCV.

Estes resultados e os anteriores, obtidos na **Figura 17**, provaram, assim, a validade do modelo e a sua possível expansão para registos não classificados. Deste modo, o modelo proposto possibilita que utilizadores que desconheçam o seu estado possam aferir o seu índice, de uma forma eficiente. No entanto, as suas principais vulnerabilidades prendem-se com o facto de o modelo:

1. Ser mais indicado para indivíduos adultos entre os 30 e os 65 anos, uma vez que existiram poucos registos utilizados no processo de DM fora deste intervalo etário.
2. Não ter em conta mais fatores de risco que, segundo a literatura, estão relacionados com o surgimento de DCV, como é o caso da etnia.
3. Ter apenas em consideração, no processo de DM, valores pontuais médios de pressão arterial alta e baixa e de dados de análises, negligenciando-se o seu historial.
4. Ser específico para população da mesma região geográfica e para registos contemporâneos.

A próxima fase consistiu em rever-se todos os processos anteriores e, uma vez consolidados e aprovados, planeou-se uma estratégia de implementação dos resultados obtidos.

3.6 Implementação

O modelo desenvolvido pode ser usado como suporte ao cálculo de índices de bem-estar cardíaco para quaisquer utilizadores que pretendam conhecer o seu valor e, como exemplo da sua aplicabilidade prática, na **Figura 19** sugere-se uma possível implementação.

Como possibilidade de utilização futura, o modelo proposto admite a implementação de aplicações móveis, que permitam que os utilizadores se registem e introduzam as suas informações pessoais e clínicas necessárias para a estimativa do seu valor de índice. De notar que as informações pessoais apenas necessitariam de ser inseridas uma única vez, e que as informações clínicas, que dizem respeito a resultados de análises, poderiam ser atualizadas de forma regular. Além disso, estas aplicações poderiam conectar-se a *smartbands*, que mediriam os valores médios diários relativos às pressões arteriais alta e baixa, e ao nível de exercício físico praticado. Assim, tendo em conta estes valores, os índices dos indivíduos poderiam ser calculados e monitorizados de forma diária. Além disso, de forma a que o historial dos utilizadores não se perdesse, poderia ser utilizado um *data warehouse* para armazenar todos os registos históricos. Deste modo, para um mesmo utilizador, as medições não necessitariam de ser calculadas de forma independente umas das outras, e o valor do índice diário retornado

poderia ser ajustado de acordo com os valores históricos do utilizador. Outras das vantagens de se armazenarem os registos na forma de um DW seria o facto de um repositório de dados facilitar as consultas e a visualização gráfica dos valores dos índices para um utilizador específico ao longo do tempo e, também, possibilitar a observação dos índices médios, tendo-se em conta o total de utilizadores, por distrito, por exemplo. Com isto, a incorporação dos índices num SDW resultaria, não só em vantagens para os utilizadores registados, como também para profissionais ligados ao setor da Saúde.

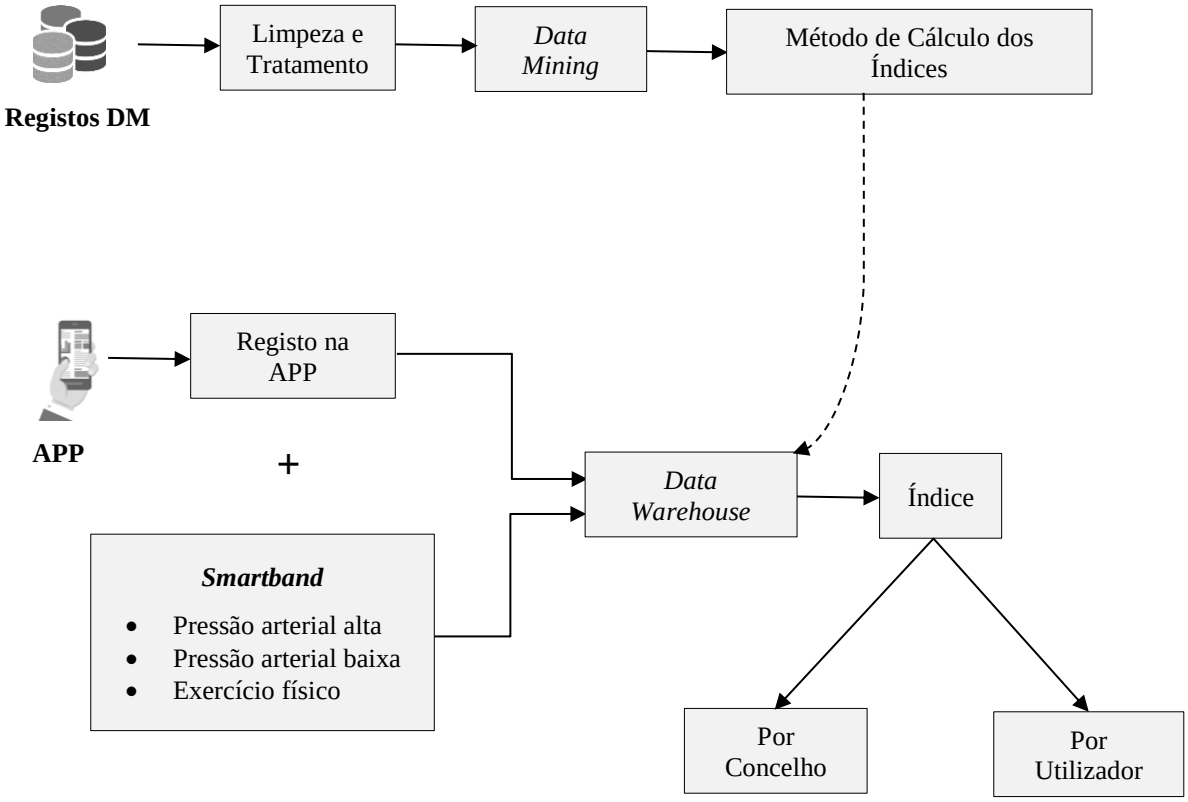


Figura 19 - Possível implementação do modelo desenvolvido.

Uma vez definida uma possível estratégia de implementação do modelo proposto, elaborou-se o seu plano de manutenção e de monitorização. Para se garantir o correto funcionamento do sistema, considerou-se essencial efetuarem-se avaliações periódicas por parte de especialistas, tendo esta avaliação uma importância acrescida pelo facto de o sistema poder funcionar de uma forma automatizada. Deste modo, os analistas de dados devem agir no sentido de averiguarem a existência de falhas, que possam condicionar o seu desempenho geral. Para tal, devem verificar, por exemplo, se existem enviesamentos ao nível dos atributos que

provoquem desequilíbrios nos dados ou se existem registos desatualizados. Assim, devem procurar responder às questões:

1. Os registos são representativos da população-alvo?
2. Os dados estão balanceados?
3. Os registos mantêm-se atuais?

De acordo com a resposta às questões colocadas, o analista, sempre que se justifique, deverá adaptar o sistema implementado.

4. O Processo de Armazenamento dos Dados

Tal como descrito, a existência de um repositório capaz de lidar eficientemente com uma grande cardinalidade de dados pode resultar num favorável complemento ao sistema de cálculo de índices desenvolvido. Como o objetivo da presente dissertação se prende com a construção de um modelo que permita, em simultâneo, uma análise regular dos valores dos índices dos utilizadores e que funcione também como uma ferramenta de suporte à decisão útil para profissionais ligados à Saúde, considerou-se vantajoso armazenar os dados na forma de um SDW.

4.1 Planeamento e Gestão do SDW

À semelhança de qualquer outro projeto, a implementação de um SDW requer um processo de gestão minucioso, ao nível do planeamento dos diversos passos a serem executados. Assim, numa primeira fase, foi necessário identificar-se os recursos disponíveis, de modo a garantir-se a satisfação de todas as necessidades inerentes à execução do SDW. Posteriormente, traçou-se um plano de desenvolvimento para a implementação do sistema, de forma a encadearem-se temporalmente as etapas necessárias, para que a implantação fosse efetuada com sucesso e no menor intervalo de tempo possível. Numa fase seguinte, foram definidas as métricas de avaliação do grau de desempenho do DW construído.

Relativamente aos recursos, para que se pudesse construir o DW, considerou-se necessária a existência de fontes de dados e a utilização do *software* Spoon e SQL Server. De notar que se constatou que, numa fase posterior à implementação propriamente dita do sistema, para que se possibilitasse a consulta e a análise pretendidas dos dados, seria preciso recorrer ao *software* Visual Studio, Power BI e QGIS. Além disso, em termos de recursos humanos, verificou-se ser precisa a presença de um *business driver*, responsável pela gestão e pela condução do projeto, e ainda de um analista, para efetuar manutenções periódicas.

Para o plano temporal do processo de implementação, definiram-se etapas, enunciaram-se precedências, discriminaram-se tarefas e atribuíram-se tempos para as executar. As principais fases consideradas foram as que diziam respeito ao planeamento do projeto, aos requisitos, à arquitetura do sistema, à implementação propriamente dita do SDW e, por último, ao seu teste e validação. Em alguns casos, as fases consideradas pressupunham ainda a execução de um conjunto específico de tarefas. Cada uma das fases e tarefas definidas foram incorporadas no diagrama de Gantt (**Figura 20**), com horizontes temporais específicos para a sua realização.

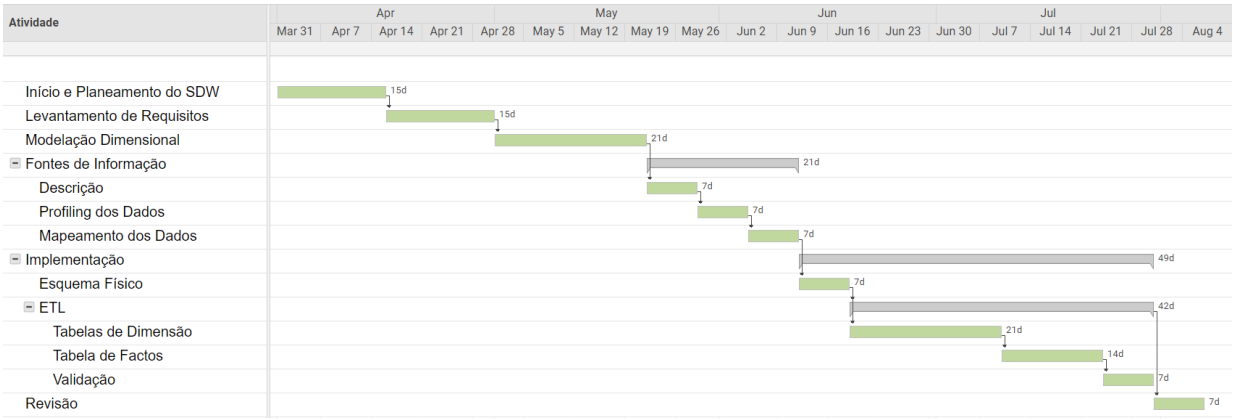


Figura 20 - Diagrama de Gantt do planeamento do projeto de SDW.

Quanto à definição dos critérios de sucesso da implementação do sistema, teve-se em conta a vertente intrínseca ao próprio DW e, sobretudo, os ganhos que advinham da sua utilização. Assim, de modo a apurar-se se o sistema construído respondia positivamente às necessidades identificadas, foi estabelecido um conjunto de métricas com o objetivo de se avaliar o seu grau de sucesso. No caso do presente trabalho, considerou-se que as métricas que melhor refletiam a qualidade do DW implementado eram as seguintes:

- O período de inatividade diário, para a execução do processo de extração, transformação e carregamento (ETL) dos dados, deverá ser inferior a 1h.
- A frequência de complicações no sistema, que requeiram intervenções por parte de especialistas externos, deverá ser inferior a três vezes por ano.
- Após um ano de utilização regular, o valor médio dos índices dos utilizadores deverá subir, pelo menos, 5%.
- O grau de satisfação dos utilizadores deverá ser superior a 90%.

4.2 Levantamentos dos Requisitos

Após o planeamento inicial, efetuou-se um levantamento dos requisitos, de modo a identificarem-se todas as necessidades inerentes à implementação do sistema. Esta etapa é fundamental para o sucesso de qualquer SDW, uma vez que, caso um dos requisitos falhe, a sua viabilidade é comprometida. Assim, antes da sua concretização, deve existir sempre um rigoroso e detalhado processo de recolha de todos os requisitos. Deste modo, para o caso em estudo, foram identificados e organizados os requisitos necessários, de acordo com a sua tipologia (descrição, exploração e controlo e acesso).

Em relação aos requisitos de descrição, que identificam os elementos fundamentais para a construção do SDW, consideraram-se os que se enumeram:

- Todos os utilizadores devem estar registados na base de dados correspondente.
- Os utilizadores devem fornecer todos os dados pessoais solicitados, assim como os dados relativos às análises e à restante informação clínica.
- Todos os distritos do país devem constar nas fontes de informação, com os campos que lhes estão associados devidamente preenchidos.
- É necessária a existência de um modelo de cálculo aplicável aos registos dos utilizadores, que permita a determinação dos seus índices.

Qualquer SDW é concebido para suprir as necessidades de se organizarem e visualizarem os dados existentes. Os dados, por si só, não transmitem informações relevantes e, para que tal aconteça, precisam de ser agrupados e mostrados de forma a realçarem-se as suas relações. A título de exemplo, ao analisar-se, de forma isolada, todos os valores dos índices dos utilizadores, pouca informação se obtém. Ao invés, se esses dados forem agrupados e os valores forem mostrados, por exemplo, pelos índices médios por região, torna-se mais fácil a identificação das localidades onde se deve investir mais em campanhas de sensibilização para a Saúde. Perceber essas necessidades de informação e transformá-las em requisitos é essencial, dado que o sucesso de um SDW depende, em última análise, de acrescentar valor aos dados nele inseridos. Desta forma, contextualizando para o que se pretendia com a implementação de um SDW para o caso em estudo, foram definidos os requisitos de exploração que se listam, relacionados com o que se espera obter da sua utilização:

- Assegurar a consulta dos índices relativos a utilizadores individuais, de forma ponderada com o seu historial.
- Observar a evolução dos índices ao longo do tempo.
- Visualizar todas as possíveis combinações dos índices, de acordo com a data, a localização e/ou o tipo de utilizador.

No grupo dos requisitos de controlo e de acesso, foram identificados os utilizadores autorizados a aceder ao DW, bem como a tipologia de consultas que cada um poderia efetuar. Estes requisitos especificaram-se do seguinte modo:

- Apenas os utilizadores registados podem consultar o DW.
- Um utilizador não pode consultar os dados relativos aos valores dos índices de outros indivíduos.
- O acesso aos dados gerais, agrupados por data ou por localização, poderá ser autorizado a qualquer utilizador.

4.3 Modelação Dimensional

Determinados os requisitos do sistema, seguiu-se a fase de modelação dimensional dos dados, que é uma das mais importantes etapas no processo de implementação de um SDW. Nesta fase é preciso definir, caracterizar e organizar as estruturas de dados a implementar, tendo em vista as perspetivas de análise pretendidas.

No que diz respeito à configuração do modelo, optou-se por se interligar a tabela de factos a várias tabelas de dimensão, num formato tipo estrela, para se possibilitar que a informação contida no DW fosse apresentada de uma forma rápida e facilmente entendível pelos utilizadores.

O processo de modelação dimensional, para o presente caso, considerou a abordagem dos quatro passos, proposta por Kimball e, como tal, foram tomadas quatro medidas principais acerca da forma de estruturação da arquitetura do DW:

1. Selecionou-se o processo de negócio.
2. Definiu-se o grão.
3. Identificaram-se as dimensões.
4. Identificaram-se os factos.

No contexto em estudo, a principal atividade do negócio relaciona-se com a realização de medições regulares ao bem-estar cardíaco da população. Neste caso, estas medições são determinadas a partir de processos de DM, suportados pelos registos dos principais fatores de risco enunciados pela literatura, que resultam em índices de bem-estar cardíaco. Como o processo de negócio está relacionado com as atividades a incluir no DW, as métricas da tabela de factos do modelo dimensional pressupõem os fatores de risco considerados no processo de cálculo, além do valor do índice resultante.

A segunda etapa da metodologia de Kimball relaciona-se com a definição do grão, ou seja, com o nível mais atómico com que são armazenados os dados no modelo dimensional. Este passo, que está intrinsecamente relacionado com as linhas da tabela de factos, é uma etapa crucial, e a mais importante no processo de modelação dimensional, que exige uma decisão ponderada e cuidada, uma vez que todas as consultas que se pretendam efetuar devem poder ser agregadas de acordo com o grão, ou múltiplos. Assim, para se selecionar o grão mais indicado foi preciso averiguar-se o tipo de consultas de dados mais relevantes a efetuar no DW e, tendo em conta as múltiplas perspetivas pretendidas para se examinar os dados, considerou-se que o nível de detalhe máximo necessário seria o valor do índice de bem-estar cardíaco para um utilizador em particular, num dado distrito e num determinado dia.

Tendo-se estabelecido o grão, o processo de identificação das dimensões tornou-se mais simples. As tabelas de dimensão consistem em conjunto de atributos fortemente relacionados e que contêm informação a partir da qual os dados do DW podem ser filtrados e agrupados. Deste modo, para se satisfazer o nível de detalhe definido, foi preciso, além da dimensão relativa ao calendário, incluir-se mais duas dimensões, respeitantes aos utilizadores e aos distritos.

A dimensão associada ao calendário foi configurada de forma a ser composta por duas hierarquias, sendo uma delas formada pelo dia, mês, trimestre e ano, e a outra pelo dia e pela semana. Esta dimensão não tem um crescimento associado, na medida em que só necessita de ser povoada uma vez e, como tal, foi projetada para um horizonte temporal de 20 anos, que se estimou ser o período de vida do DW.

De forma similar, a dimensão que diz respeito aos distritos também apresentou um comportamento estático, uma vez que, tendo em conta o período de vida do DW, não seriam expectáveis alterações nas regiões administrativas em Portugal, que conduzissem a alterações nos seus distritos. Assim, mais uma vez, esta dimensão apenas precisou de ser povoada uma única vez e não teve um crescimento associado. O objetivo da inclusão desta dimensão foi o de se poder possibilitar a conceção de sistemas de apoio à decisão que tivessem em consideração o distrito, a província, a região e a localização litoral/interior. Neste caso, foram consideradas duas hierarquias diferenciadas entre os atributos.

Por outro lado, a dimensão relativa aos utilizadores teve o propósito de incorporar informações acerca da sua localização e sexo, mas também detalhes complementares acerca das características económicas e literárias dos indivíduos, de forma a possibilitarem-se análises futuras, tendo em conta esses parâmetros. Como, à exceção do sexo, todos os atributos dos utilizadores são suscetíveis a variações, esta dimensão apresenta variação. Além disso, como em qualquer instante se possibilita a inserção e a remoção de utilizadores, ao contrário das dimensões anteriores, esta dimensão tem um comportamento dinâmico.

Para cada uma das dimensões consideradas, foi construída uma tabela que resume as suas principais características. Estas tabelas podem ser consultadas no Anexo V.

A última etapa do processo de modelação dimensional consistiu na identificação dos factos, ou seja, das medidas importantes a registar em cada transação do negócio, que não são conhecidas à partida. Assim, os factos considerados foram, além do próprio valor de índice de bem-estar, as métricas que sustentam o seu cálculo, isto é, a idade, o sexo, o IMC, os antecedentes familiares, a quantidade de cigarros fumada, o colesterol total, a diabetes, a hipertensão, a dor após esforço, o hipotiroidismo, as pressões arteriais alta e baixa, e o exercício

físico. Analogamente às dimensões, para a tabela de factos foi também desenvolvida uma matriz, no Anexo VI, que reúne as suas principais características.

Da execução destas etapas resultou o esquema conceptual que se apresenta **Figura 21**, composto por três tabelas de dimensão e uma de factos, que serviu de suporte à implementação física do SDW.

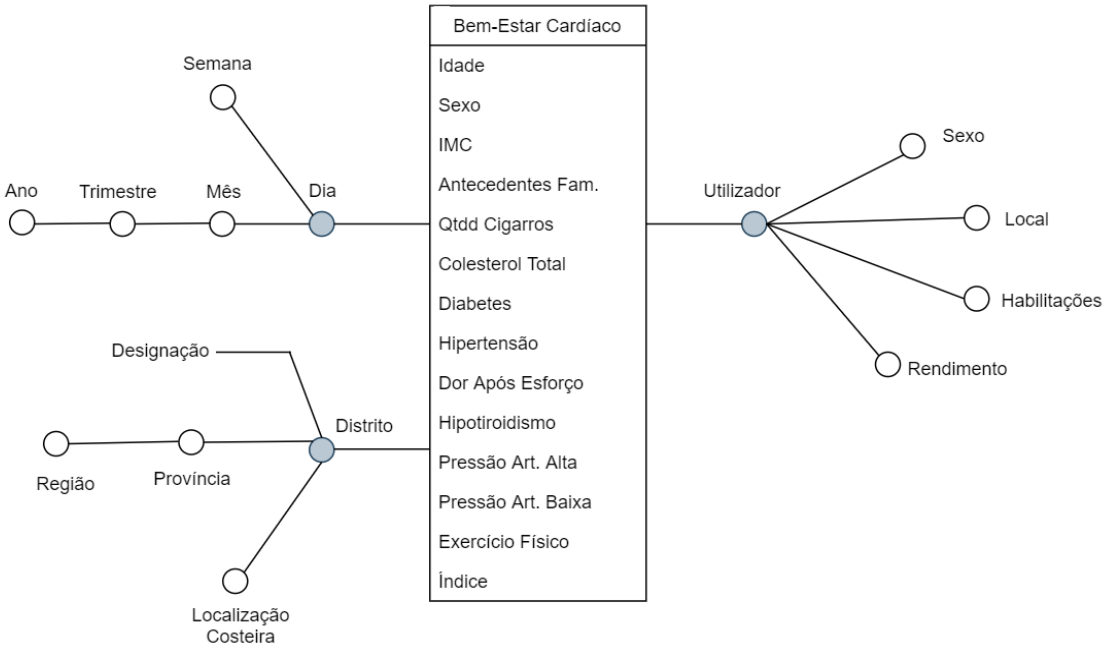


Figura 21 - Esquema conceptual do caso em estudo.

4.4 Fontes de Informação

Uma das fases essenciais para se garantir a viabilidade do sistema a implementar passa pela análise da disponibilidade dos dados nas fontes e pela verificação da sua qualidade. De facto, a satisfação dos critérios mínimos de sucesso é condição *sine qua non* para que qualquer SDW possa ser implementado com êxito. Deste modo, depois de definido o modelo conceptual, e ainda numa etapa prévia ao começo da implementação, tornou-se necessário averiguar se estes critérios eram cumpridos.

Ao nível da disponibilidade dos dados, analisaram-se as fontes candidatas e a sua composição. Estas fontes, para efeitos de identificação, foram designadas por Fonte 1, Fonte 2 e Fonte 3 e foram sintetizadas especificamente para a verificação do sistema proposto. Do ponto de vista descritivo, as suas principais origens e características são as que se listam na **Tabela 13**. Desta forma, estas fontes dizem respeito aos utilizadores registados numa APP, e que nela tenham inserido os seus dados pessoais e clínicos. Os dados pessoais não são editáveis, ao passo que os dados clínicos podem ser atualizados na fonte sempre que estes façam novas análises

médicas. Além disso, estes utilizadores podem associar um equipamento eletrónico, conectado à Fonte 2, que permita estimar, diariamente, os seus valores médios das pressões arteriais e as calorias gastas.

Tabela 13 - Sumário das fontes candidatas

Fonte	Formato	Descrição
Fonte 1	Base de dados	Compila os dados provenientes de uma APP relativos aos registos pessoais e clínicos de vários utilizadores.
Fonte 2	Base de dados	Recorre a um servidor central, que recebe dados relativos às pressões arteriais alta e baixa, e às calorias despendidas. Estes dados são recolhidos de vários utilizadores através de dispositivos móveis do tipo *smartband* e referem-se aos seus valores médios diários.
Fonte 3	Ficheiro CSV	Reúne informação geográfica relativa aos distritos do país.

Como as fontes 1 e 2 foram criadas exclusivamente para o propósito de calcular índices de bem-estar cardíaco para os seus utilizadores, garantiu-se que todos os parâmetros necessários para a sua determinação eram solicitados, além das informações pessoais acerca do seu rendimento, habilitações literárias e local de residência. Nos casos em que seria provável que os utilizadores não soubessem precisar valores, como o IMC ou a quantidade de cigarros fumada, optou-se por se requerer parâmetros semelhantes, mas de mais simples resposta que, a partir de fórmulas, pudessem determinar as métricas pretendidas para o cálculo dos índices. Por sua vez, para que os utilizadores não tenham de atualizar o seu registo sempre que completem mais um ano de vida, foi-lhes pedida a data de nascimento, em detrimento da idade, para que, a partir dela, se possa calcular, a idade correspondente a cada transação. Assim, no que diz respeito ao povoamento da dimensão *Utilizador* e da tabela de factos *Bem-Estar Cardíaco*, assegurou-se que, no seu conjunto, as fontes 1 e 2 forneciam todos os campos necessários à sua implementação no DW. De forma complementar, avaliou-se o conteúdo da Fonte 3 e verificou-se que era suficiente para carregar a tabela associada à dimensão *Distrito*.

De uma forma mais pormenorizada, a **Tabela 14** discrimina e descreve todos os parâmetros constituintes de cada uma das fontes, e evidencia a estrutura do modelo dimensional em que esses parâmetros poderão ser utilizados. Da tabela verifica-se que nenhuma das fontes pode ser suprimida, sendo todas elas necessárias para a implementação do SDW.

Tabela 14 - Conteúdo das fontes candidatas

Fonte	Atributo	Descrição	Dim Utilizador	Dim Distrito	Tabela de Factos
Fonte 1	*id* utilizador	Identificação do utilizador	x		x
	data análises	Data das análises a que os exames médicos dizem respeito			x
	data nascimento	Data de nascimento do utilizador			x
	sexo	Género do utilizador	x		x
	distrito	Local de residência do utilizador	x		
	habilitacoes	Habilitações literárias do utilizador	x		
	rendimento	Rendimento do utilizador, em milhares	x		
	peso	Peso do utilizador, em kg			x
	altura	Altura do utilizador, em cm			x
	antecedentes	Histórico familiar de DCV			x
	fumador	Indica se o utilizador foi ou é fumador			x
	num_cigarros	Número médio de cigarros fumado por dia			x
	num_anos_fum	Número de anos que fumou/fuma			x
	colesterol_total	Valor do colesterol total, em mg/dL			x
	glicose_rapida	Valor da glicose rápida, em mg/dL			x
	diabetes	Indica se o utilizador é diabético			x
	hipertensao	Indica se o utilizador é hipertenso			x
	dor_esforco	Indica se o utilizador sente dor após praticar esforços físicos			x
	hipotiroidismo	Indica se o utilizador tem hipotiroidismo			x
Fonte 2	*id*	Identificação do utilizador	x		x
	data_índice	Data em que os registos diários foram medidos			x
	data_registo	Data e hora em que os registos foram inseridos na fonte de dados			x
	calorias	Número de calorias gastas, num dia, pelo utilizador			x
	pa_alta	Valor médio diário da pressão arterial alta			x
	pa_baixa	Valor médio diário da pressão arterial baixa			x
Fonte 3	codigo	Identificação do distrito		x	
	nome	Nome do distrito		x	
	província	Província do distrito		x	
	região	Região do distrito		x	
	litoralinterior	Indica se o registo se situa no litoral ou no interior do país		x	

A garantia de que todos os campos necessários estão contidos no conjunto de fontes de dados considerado não é suficiente para que se conclua acerca da qualidade dessas fontes. Nesse sentido, foi efetuado um processo de *profiling* dos dados para as fontes em estudo, ou seja, foram gerados perfis de dados, que visaram avaliar as suas principais características e estatísticas. Estes perfis são capazes de fornecer informação acerca dos valores omissos (ou nulos), dos valores únicos, do intervalo de variação e distribuição das variáveis, e possibilitam ainda análises ao seu formato e tamanho (Rodrigues, Coles and Dye, 2012).

Como as fontes 1 e 2 correspondiam a bases de dados do SQL Server e é possível realizar-se o processo de *profiling* através da ferramenta *SQL Server Data Tools*, recorrendo-se ao *Data Profiling Task Editor* do Visual Studio, esta análise foi efetuada a partir deste instrumento, para estas duas fontes. Assim, foram selecionados vários tipos de perfis, que permitiram avaliar e verificar os dados, com base em diferentes perspetivas:

- ***Candidate Key Profile Request*** – Permitiu testar a adequabilidade de se usar colunas específicas como chave, indicando potenciais problemas, como a existência de valores duplicados.
- ***Column Length Distribution Profile Request*** – Identificou todos os comprimentos distintos existentes em determinadas colunas do tipo *string*, bem como a sua percentagem relativamente à totalidade da tabela.
- ***Column Null Ratio Profile Request*** – Indicou a percentagem de valores nulos existentes nas colunas selecionadas.
- ***Column Pattern Profile Request*** – Informou acerca da percentagem da frequência de utilização de cada uma das expressões regulares dos campos nominais.
- ***Column Statistics Profile Request*** – Reportou valores de estatísticas básicas, tais como a média, o desvio-padrão e os valores mínimos e máximos dos campos numéricos e das datas.
- ***Column Value Distribution Profile Request*** – Relatou acerca de todas os valores distintos presentes em colunas específicas, bem como a sua percentagem em relação à totalidade dos registos.
- ***Functional Dependency Profile Request*** – Assinalou o grau de dependência dos valores de uma coluna, relativamente aos de outra.
- ***Value Inclusion Profile Request*** – Determinou a sobreposição de valores entre colunas distintas e permitiu averiguar se uma delas poderia funcionar como chave estrangeira da outra.

Depois de implementados os tipos de perfis selecionados para a análise de ambas as fontes, obteve-se a arquitetura que se visualiza na **Figura 22** e, através da opção *Open Profile Viewer*, foi possível observarem-se os resultados, reproduzidos no Anexo VII.

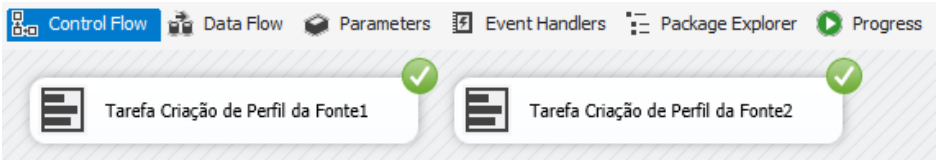


Figura 22 - Criação das tarefas de *profiling* dos dados das Fontes 1 e 2.

Relativamente à Fonte 1, os perfis revelaram que a coluna *id* era uma chave apropriada, uma vez que 100% dos seus valores não eram duplicados, nem se verificou qualquer outro entrave, que a inviabilizasse de ser usada como chave primária. Além disso, os resultados obtidos também permitiram constatar que os valores inseridos nas diversas colunas tinham um comprimento compatível com aquele que seria expectável. A título de exemplo, as colunas nominais que apenas podem conter as expressões "sim" ou "não" apresentaram, quer para o comprimento mínimo, quer para o máximo, o valor "3". Outro exemplo ilustrativo é o que se refere ao campo das habilitações literárias, em que se verificou que 2004 registos tinham um comprimento de 12 carateres (e, por isso, em princípio, corresponderiam às *strings* "licenciatura" ou "doutoramento"), 1036 um comprimento de 6 carateres (compatível com "básico"), 1045 instâncias tinham 10 dígitos (que estão de acordo com "secundário") e existiam ainda 957 registos cujos valores eram constituídos por 8 letras (que poderiam, assim, advir do nome "mestrado"). Posto isto, verificou-se que a quantidade de valores distintos, neste caso "5", estava em sintonia com os registos permitidos. Como outro exemplo, no caso do id, existiam 5042 valores únicos, ou seja, tal como já se tinha provado, todos os registos tinham um *id* distinto, uma vez que esta fonte dispõe de um total de 5042 registos. Além disso, quanto às colunas em que apenas é aceitável um valor do tipo "sim" ou "não", observou-se que todas elas apresentavam, na sua totalidade, dois tipos de valores distintos. Desta forma, estendendo a linha de raciocínio para os restantes dados, confirmou-se que todos os seus valores eram compatíveis com o previsto. No que toca à presença de dados omissos, comprovou-se que todas as colunas estavam integralmente preenchidas. Ao nível das estatísticas básicas obteve-se o quadro resumo indicado na **Figura 23**, que permitiu notar que, de uma forma geral, os valores mínimos e máximos se enquadravam relativamente bem com aquilo que seria esperado. Por exemplo, a altura variou entre 150 e 197 cm e as datas de análises estavam contidas no intervalo entre os dias 22/09/2017 e 23/08/2019. Em termos de valores médios, os resultados também foram coerentes e, no que se refere ao desvio-padrão, os valores retornados foram relativamente

baixos em comparação ao valor médio de cada coluna. Neste caso, a única exceção foi verificada no rendimento, cujo desvio-padrão mostrou uma ordem de grandeza elevada. Obviamente, importa realçar que, no caso da coluna *id*, a média e o desvio-padrão não têm qualquer significado tangível.

Coluna	Mínimo	Máximo	Média	Desvio Padrão
altura	150	197	175.576953589...	11.5569630905385
colesterol_total	70	230	147.909758032...	28.8682392258096
data_analises	01/01/2017 00:...	31/01/2019 00:...		
data_nascimento	01/01/1949 00:...	30/01/1989 00:...		
glicose_rapida	65	170	90.0955969853...	22.12054495204
id	1	5042	2521.5	1455.5
num_anos_fum	0	35	7.49464498214...	10.5903524182355
num_cigarros	0	40	10.2578341927...	14.1748083409006
peso	46	175	97.3843712812...	23.6620187680821
rendimento	0	70	34.750892502975	20.4859697007071

Figura 23 - Perfis de estatísticas por coluna, referentes à Fonte 1.

Assim, em relação a esta fonte, de uma forma geral, verificou-se que os dados nela contidos tinham qualidade.

No que respeita à Fonte 2, o número de valores distintos em cada coluna aparentou ser adequado. No entanto, relativamente ao *id* verificaram-se mais dois valores distintos do que os da Fonte 1, o que indiciou que, pelo menos, mais dois utilizadores tinham transações na Fonte 2, sem estarem registados na Fonte 1. O perfil de dados referente ao padrão por coluna possibilitou um nível maior de detalhe para se analisar esta situação. Efetivamente, comprovou-se que os *ids* 6000 e 7000, da Fonte 2, não tinham qualquer correspondência com os *ids* da Fonte 1. Assim, globalmente, os *ids* da Fonte 2 tiveram uma correspondência de 99.998% com os da Fonte 1 e, por isso, é possível que estes atributos funcionem como chave de interligação entre as tabelas respetivas. No entanto, durante o processo de tratamento de dados, os registos com *ids* que existam na Fonte 2, mas que não existam na Fonte 1, deverão ser descartados. Em relação aos valores nulos, à semelhança da fonte anterior, verificou-se que todas as colunas estavam devidamente preenchidas. No que se refere às estatísticas básicas, a **Figura 24** mostra que os valores associados a cada um dos campos estavam de acordo com aquilo que seria razoável esperar.

Coluna	Mínimo	Máximo	Média	Desvio Padrão
calorias	50	2655	1175.14033	649.601838419113
data_índice	02/01/2017 00:...	05/10/2019 00:...		
data_registo	10/01/2017 00:...	15/10/2019 00:...		
Id	1	7000	2400.28459	1493.21319512604
pa_alta	100	167	118.862	14.5905632516363
pa_baixa	65	116	80.22261	10.1527826130525

Figura 24 – Perfis de estatísticas por coluna, referentes à Fonte 2.

Deste modo, também se constatou que a Fonte 2 apresentava a qualidade mínima exigida para ser considerada no processo de povoamento do DW.

Ao contrário das fontes anteriores, o processo de verificação da qualidade dos dados da Fonte 3 não foi efetuado com recurso a ferramentas informáticas, uma vez que esta fonte apresentava uma baixa cardinalidade, dispondo apenas de um total de 20 registos e de 5 parâmetros. Assim, neste caso em particular, o processo de avaliação foi passível de ser efetuado de forma manual, tendo-se observado que todos os campos estavam devidamente preenchidos, sem valores omissos e dentro do expectável, e que o código de cada distrito era único.

Desta forma, conseguiu-se assegurar a disponibilidade dos dados e a qualidade de cada umas das fontes consideradas. No entanto, como nem todos os dados têm ainda uma correspondência direta entre a origem e o destino, foi elaborada a estrutura de mapeamento da **Tabela 15**, que mostra as transformações necessárias efetuar, para tornar os dados das fontes aptos a serem inseridos no DW.

Tabela 15 - Mapeamento entre as fontes de dados e o DW

Origem	Atributo	Transformação			Atributo	Destino
Fonte 1	id	int	→	int	id_utilizador	Dim Utilizador
	distrito	varchar	→	varchar	distrito	
	habilitacoes	varchar	→	varchar	habilitacoes	
	rendimento	int	→	varchar	rendimento	
	sexo	varchar	→	varchar	sexo	
		varchar	→	varchar	sexo	Tabela de Factos
	peso	int	Transf.1 →	float	IMC	
	altura	int				
	antecedentes	varchar	→	varchar	antecedentes	
	colesterol_total	int	→	int	colesterol_total	
	glicose_rapida	int	→	int	glicose_rapida	
	diabetes	varchar	→	varchar	diabetes	
	hipertensao	varchar	→	varchar	hipertensao	
	dor_esforco	varchar	→	varchar	dor_esforco	
	hipotiroidismo	varchar	→	varchar	hipotiroidismo	
	data_analises	date	Transf.2 →	int	idade	
	data_nascimento	date				
	num_cigarros	int	Transf.3 →	float	qtdd_fumador	
	num_anos_fum	int				
Fonte 2	id	int	→	int	id_utilizador	Tabela de Factos
	data_indice	date	→	date	data_indice	
	calorias	int	→	varchar	exercicio_fisico	
	pa_alta	int	→	int	pa_alta	
	pa_baixa	int	→	int	pa_baixa	
Fonte 3	código	texto	→	varchar	código	Dim Distrito
	nome	texto	→	varchar	nome	
	provincia	texto	→	varchar	provincia	
	região	texto	→	varchar	região	
	litoralinterior	texto	→	varchar	litoralinterior	

Legenda:

Transf.1 $IMC = \frac{peso}{altura^2}$

Transf.2 idade $= Anos(data\_analises - data\_nascimento)$

Tranf.3 $qtdd\_Fumador = \frac{365 \times num\_cigarros \times num\_anos\_fum}{1000}$

4.5 Implementação do SDW

Tendo-se constatado a adequabilidade das fontes de informação descritas, foi preciso ter em consideração o seu relacionamento com as restantes áreas a incluir no SDW. Além do Sistema de Informação (SI), que contém as fontes de informação, é essencial a existência de uma Área de Retenção (DSA), além do próprio DW. Desta forma, o SDW implementado foi suportado em três áreas distintas, conforme se visualiza na **Figura 25**. Assim, o mecanismo de migração dos dados consistiu na extração dos dados do SI para a DSA, para poderem ser devidamente limpos e processados, de modo a estarem preparados a serem incluídos no DW.

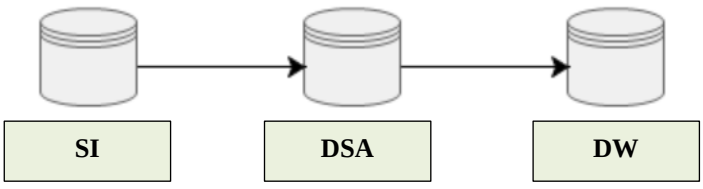


Figura 25 - Mecanismo de migração dos dados.

4.5.1 Implementação dos Esquemas Físicos

Antes de se proceder à implementação propriamente dita do SDW, foi preciso criar-se as estruturas físicas necessárias para o tratamento e para o armazenamento dos dados. Deste modo, na DSA, foram produzidas as tabelas que se mostram na **Figura 26**, de forma a possibilitarem a limpeza e a manipulação eficientes dos dados.

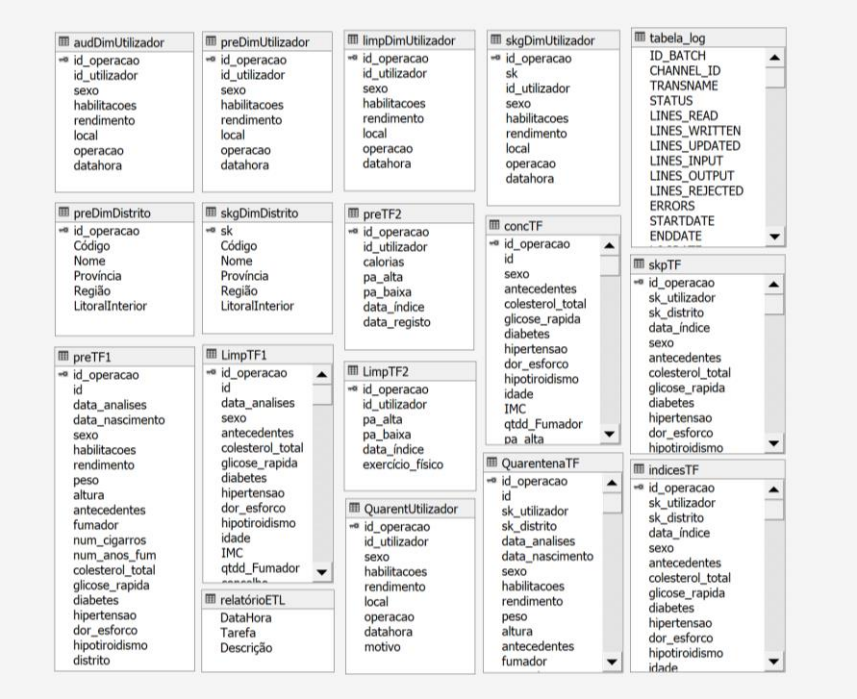


Figura 26 - Estruturas implementadas na DSA.

Para o DW foram construídas as tabelas de dimensão e a tabela de factos descritas e, para além destas, foi ainda elaborada a tabela relativa ao histórico do utilizador, de modo a arquivar, com indicação do registo temporal, todas as alterações detetadas nos valores dos seus atributos. Deste modo, o seu esquema físico segue a estrutura representada na **Figura 27**.

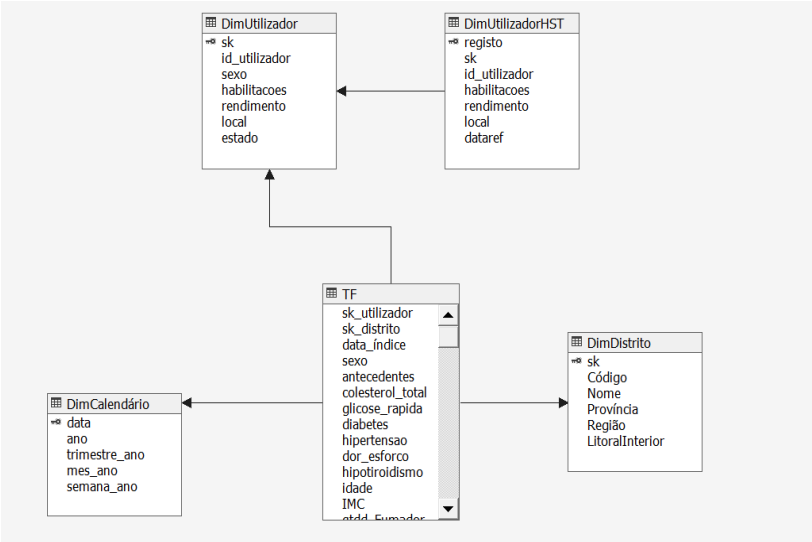


Figura 27 - Estruturas implementadas no DW.

4.5.2 Aspetos Gerais da Implementação

Uma vez criadas as estruturas físicas, que serviram de esqueleto ao processo de implementação, foi preciso definir-se a metodologia mais adequada para o carregamento das dimensões e da tabela de factos no DW.

No seu sentido mais abstrato, a concretização do DW pressupõe uma fase relativa ao processo de DM, para a seleção do algoritmo mais adequado para servir de base ao cálculo dos índices a integrar na tabela de factos, e uma outra etapa responsável pelo processo de ETL dos dados. Estas etapas podem ser efetuadas de forma assíncrona, na medida em que o processo de DM não necessita de ser efetuado com uma periodicidade tão frequente como o do ETL. Desta forma, no Spoon, foram criadas duas *jobs* principais de alto nível, que possibilitam que todo o processo de DM e ETL seja, respetivamente, desencadeado através da sua execução. Estas etapas foram separadas em *jobs* distintas, de forma a poderem ser iniciadas em diferentes períodos de tempo. No caso do processo de DM considerou-se suficiente uma execução pontual, realizada mensalmente, uma vez que, à partida, o *dataset* que suporta a construção do modelo de DM mais adequado não deve apresentar grandes variações diárias, que conduzam à utilização de um algoritmo diferente. Assim, programou-se o início da *job* referente ao DM de acordo com os parâmetros indicados na **Figura 28**, de modo a ser executada mensalmente, no primeiro dia de cada mês, com um começo às 7:00. Por outro lado, para a execução da *job* relativa ao ETL, definiu-se uma periodicidade diária, com início às 00:00.

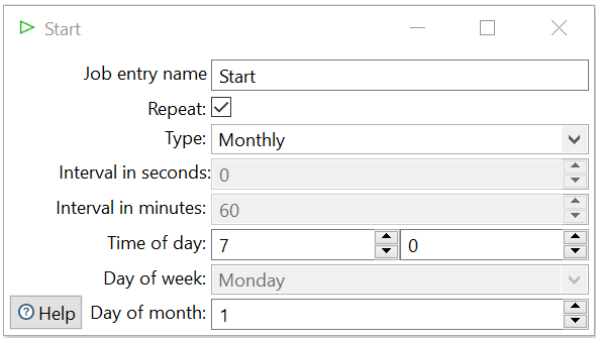


Figura 28 - Programação do início de execução da *job* relativa ao DM.

As duas *jobs* principais implementadas para o desencadeamento dos processos de DM e ETL são as que se representam nas **Figura 29** e **Figura 30**, por esta ordem. No caso da *job DM*, é preciso notar que a sua execução apenas faz sentido quando o *dataset* que contém os dados classificados para suporte ao DM inclui um grande número de registos. Caso contrário, caso haja registos em número insuficente, os algoritmos de cálculo podem não contemplar todos os parâmetros necessários e revelarem-se inadequados para a determinação dos índices. Deste

modo, optou-se por se permitir apenas efetuar este processo caso haja, pelo menos, 5000 registos no *dataset* referente ao DM. Assim, caso não se verifiquem registos em número suficiente para que se consiga apurar uma técnica de cálculo, é enviada uma mensagem de *email* ao analista, para o informar da situação. Num cenário como este, o processo fica diferido e apenas é concretizado quando a *job* for de novo executada e o número de registos for superior, ou igual, a 5000. Para a contagem do número de linhas do *dataset* relativo ao DM, foi criada uma transformação que soma o seu número de linhas e o converte numa variável designada por "num_linhas". Assim, o processo de DM apenas é realizado quando o valor desta variável é superior (ou igual) a 5000.

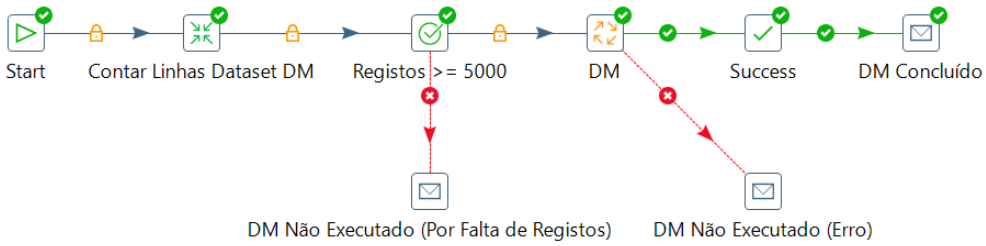


Figura 29 - *Job* principal responsável pelo desencadeamento do processo de DM.

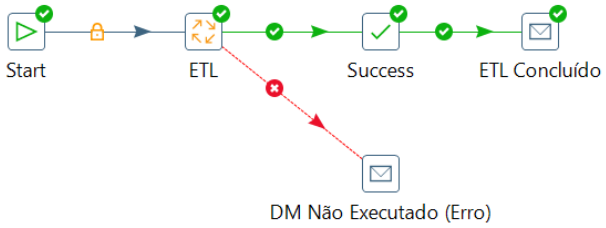


Figura 30 - *Job* principal responsável pelo desencadeamento do processo de ETL.

A *job* associada ao processo de DM corresponde às etapas de *Preparação de Dados* e de *Modelação* do processo CRISP-DM, descritas no capítulo anterior. Deste modo, esta *job* foi composta pela transformação detalhada na **Figura 7**, que permite preparar o *dataset*, que contém os dados classificados para a aplicação dos algoritmos de DM, e integrá-lo com o Weka, de forma a apurarem-se resultados para cada uma das técnicas. Além disso, foi ainda incorporada a transformação representada na **Figura 14**, que possibilita a seleção automática da técnica mais adequada para a determinação dos índices dos utilizadores, de acordo com os resultados obtidos. A técnica é escolhida tendo em conta o maior valor das pontuações ponderadas pela acurácia e pela sensibilidade. No final desta transformação, o melhor algoritmo é anotado num ficheiro de texto para poder ser, posteriormente, utilizado como suporte ao cálculo dos índices, na *job ETL*. Deste modo, a *job* utilizada para o processo de DM foi

implementada conforme se observa na **Figura 31**, sendo constituída por uma primeira transformação, que corresponde à transformação da **Figura 7**, e pela transformação da **Figura 14**.

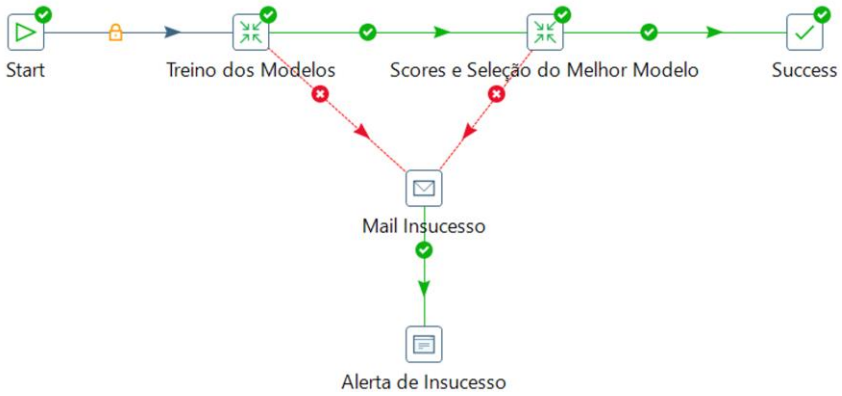


Figura 31 - *Job* relativa ao processo de DM.

Por sua vez, a *job* que diz respeito ao processo de ETL foi construída de forma a que, numa primeira fase, fossem criadas as dimensões e efetuado o tratamento dos registos a incluir na tabela de factos. Apenas depois de concluídas estas etapas é que se pôde proceder ao carregamento desta última tabela. Na **Figura 32** apresenta-se a forma de execução do processo de ETL. De salientar que, como as fontes relativas às dimensões *Calendário* e *Distrito* são estáticas, apenas devem ser povoadas uma única vez no DW e, por isso, a região assinalada a vermelho na figura diz apenas respeito ao primeiro povoamento dos dados, e não deve ser processada em povoamentos posteriores.

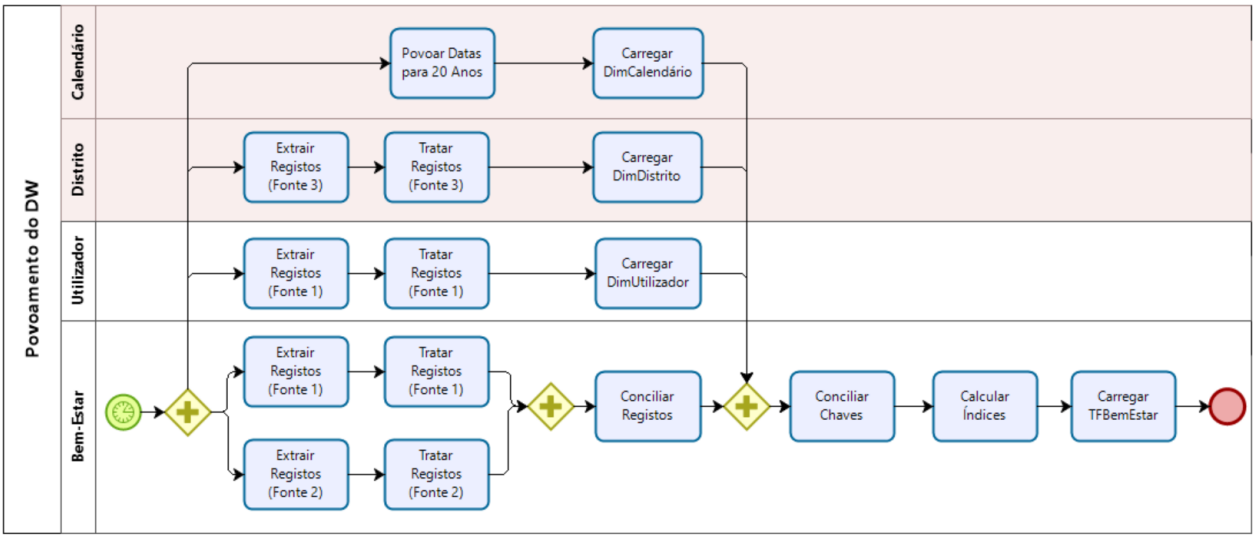


Figura 32 - Esquema BPMN da metodologia de povoamento do DW.

De forma a garantir-se este encadeamento, criaram-se duas *jobs* dentro da *job* relativa ao ETL, com o propósito de se assegurar que, em primeiro lugar, tem de ser processada a que se relaciona com o povoamento das dimensões e com o tratamento dos dados a serem integrados na tabela de factos, e só depois é que se permite a execução da job responsável pelo carregamento da tabela de factos no DW.

Deste modo, a primeira *job* inserida na *job ETL* foi implantada de acordo com a **Figura 33**. Neste caso, as *sub-jobs* que se relacionam com o carregamento das dimensões e com a preparação dos dados da tabela de factos podem ser realizadas de forma paralela. Para se assegurar que as dimensões *Distrito* e *Calendário* são apenas povoadas uma vez, foi acrescentado um contador de registos para cada uma delas. Este contador permite verificar se as dimensões estão povoadas no DW e, caso não estejam, ou seja, caso o número total de registos seja zero, as *jobs* associadas aos seus carregamentos são executadas. No entanto, caso já se tenha efetuado o povoamento destas dimensões, as *jobs* respetivas não são invocadas e ocorre de imediato "sucesso".

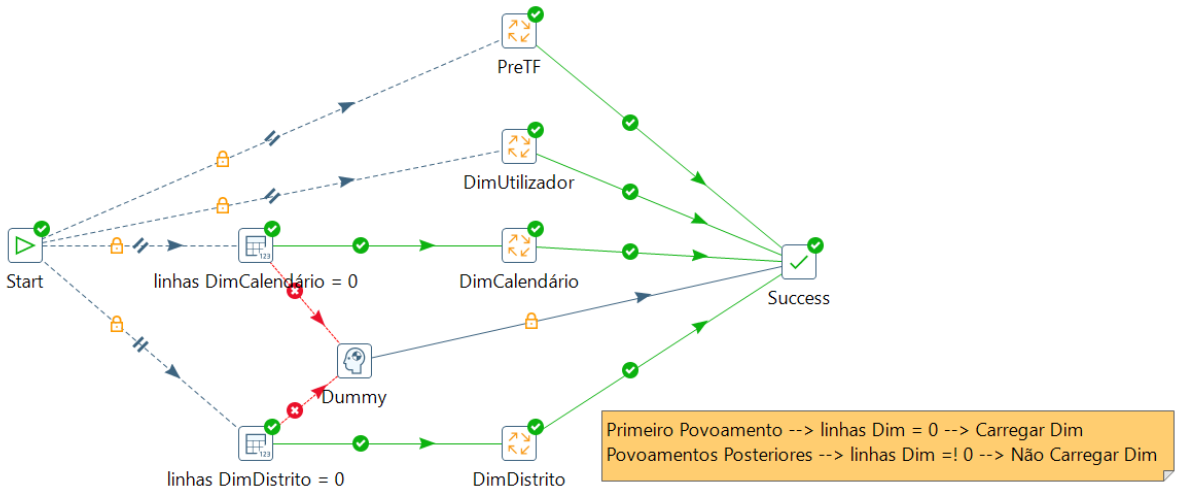


Figura 33 - Primeira *job* incorporada na *job ETL*.

Depois de efetuada, com sucesso, a *job* da **Figura 33**, é executada a *job* representada na **Figura 34**. Esta *job* trata da conciliação das chaves das dimensões com os registos a serem inseridos na tabela de factos, calcula o índice de bem-estar dos novos registos e, no final, carrega a tabela de factos no DW, terminando-se, assim, o processo de ETL.

Figura 34 – Segunda *job* incorporada na *job ETL*.

4.5.3 ETL das Tabelas de Dimensão

De um ponto de vista mais detalhado, o processo de ETL efetuado para as tabelas de dimensão e de factos exigiu um grande cuidado e um tratamento diferenciado para o povoamento de cada tabela. Este processo, que é *a priori* de fácil entendimento, é de difícil implementação prática e consome cerca de 70% dos recursos necessários para a construção do SDW (Kimball and Caserta, 2004). Deste modo, o ETL é um processo crítico e complexo, que requer esforço e tempo para que o seu planeamento seja bem estruturado. Um planeamento bem estruturado é fundamental para que se acrescente valor aos dados e se garanta sucesso na implementação do SDW. Assim, para o presente caso de estudo, para cada tabela, foi delineada uma metodologia específica de ETL para os seus dados.

Em relação às dimensões, o processo de ETL associado à dimensão *Utilizador* foi o que careceu de uma maior complexidade. Numa primeira fase, foi criada uma *job* geral (representada na **Figura 33** pela *job DimUtilizador*), cujo fluxo se pode observar na **Figura 35**, constituída por duas transformações e uma *sub-job*, relativas aos processos de extração dos dados, transformação e carregamento, respetivamente.

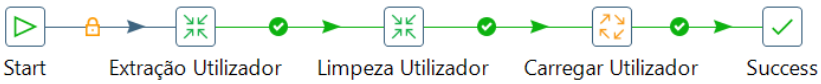


Figura 35 - Constituição da *job DimUtilizador*.

A transformação associada à extração dos dados foi a responsável pela transferência dos dados da fonte para a DSA. Importa notar que a DSA é uma área situada entre a fonte e o DW, à qual os utilizadores finais não têm acesso, e que permite efetuar todos os passos de limpeza e de conciliação necessários, desde o momento em que os dados são extraídos da fonte até ao instante em que são armazenados no DW. A principal vantagem da existência desta área intermédia é o facto de tornar o processo de ETL menos intrusivo, ao possibilitar que os utilizadores finais continuem a ter acesso às fontes e ao DW no período de tempo em que o processo de transformação dos dados está a ser executado. (Kimball and Caserta, 2004)

Para a dimensão em estudo, durante o processo de extração dos dados deve ter-se em atenção o tipo de povoamento a realizar, ou seja, deve-se averiguar se se trata de um povoamento total ou incremental. Na situação do primeiro povoamento, este procedimento é mais transparente e os dados são extraídos diretamente da fonte para a DSA, na sua totalidade. No entanto, como os SDW são formados por grandes conjuntos de dados, não é eficiente efetuar-se os povoamentos seguintes de forma integral. Deste modo, nos casos de povoamento

incremental, apenas os novos dados ou dados que tenham sido entretanto atualizados/removidos desde o último processo de ETL devem ser recolhidos e transferidos para a DSA, uma vez que os restantes registos já foram anteriormente inseridos no DW. Este procedimento de deteção exclusiva dos dados novos, atualizados ou removidos desde o último ETL designa-se por *Change Data Capture* (CDC). Existem múltiplas técnicas de CDC que podem ser aplicadas e a política de extração adotada, ou seja, a metodologia seguida para a extração dos dados das fontes, deve ter em conta a técnica mais adequada ao contexto em análise. Em relação à extração dos dados para o povoamento da dimensão *Utilizador*, considerou-se que o mecanismo CDC mais apropriado era através do uso de *triggers*. Esta solução teve um carácter intrusivo, uma vez que implicou a adição de *triggers* à fonte de dados (Fonte 1), embora este comportamento intrusivo não fosse tão acentuado como seria a manipulação direta ao conteúdo da fonte. Como a dimensão em análise apresentava variação lenta, o uso de *triggers* permitiu capturar, além do dos novos utilizadores, todos os registos removidos ou que sofreram modificações nos campos relativos às habilitações literárias, rendimento e/ou local de residência. (Casters, Bouman and Dongen, 2010) Para isso, os *triggers* implementados foram responsáveis por invocar *stored procedures* que, mediante novas inserções, alterações (associadas às habilitações, rendimento ou local de residência) ou remoções nos dados da fonte, automaticamente copiam os registos associados para uma tabela de auditoria, acrescentando-lhes duas novas colunas, referentes ao tipo de modificação (novo, atualizado ou removido) e à data e hora de captura. Os *triggers* e os *stored procedures* criados estão detalhados no Anexo VIII. Desta forma, a transformação relativa ao processo de extração dos dados, para o povoamento da *DimUtilizador*, foi criada de acordo com a metodologia apresentada na **Figura 36**. Assim, caso se trate de um povoamento total, os dados são extraídos diretamente da fonte e copiados para uma tabela na DSA, designada por *preDimUtilizador*, com indicação de que se tratam de registos novos e com a denotação da data e hora do momento em que foram copiados. Pelo lado contrário, caso o povoamento seja incremental, os registos são captados da tabela de auditoria, copiados para a tabela *preDimUtilizador* e apagados da tabela de auditoria. É importante realçar que estes passos tiveram de ser executados como uma transação, ou seja, corresponderam a uma operação única, em que, ou eram efetivados os dois passos ou, então, em caso de falha num dos processos, não era efetuada a cópia para a *preDimUtilizador* nem a auditoria era apagada, e a operação era abortada por processos de *rollback*. Desta forma, conseguiu-se garantir que, em todas as situações, a *preDimUtilizador* corresponderia a uma cópia fidedigna da tabela de auditoria. Além disso, caso surgissem registos adicionais, removidos ou modificados durante este processo na tabela-fonte, estes registos ficariam em "fila de espera" até que esta transformação

fosse concluída e, nessa altura, seriam inseridos na tabela de auditoria, para estarem disponíveis a serem utilizados na execução do próximo ciclo de ETL. Para se indicar que a transformação era transacional foi selecionada a opção "*Make the transformation database transactional*", nas suas propriedades.

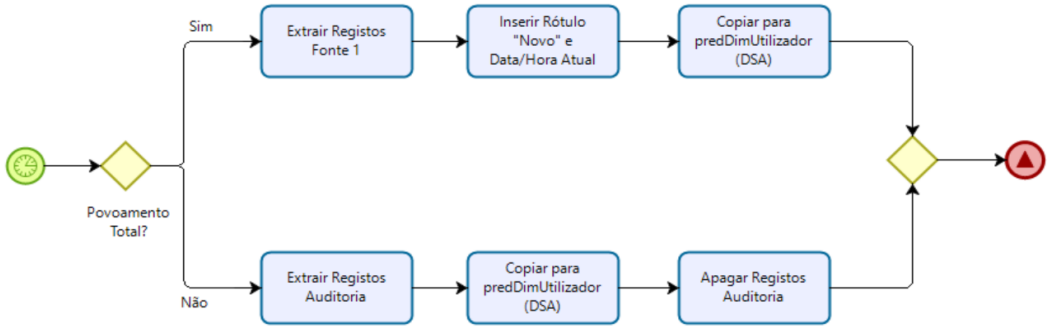


Figura 36 - Esquema BPMN da metodologia de extração dos dados para o povoamento da *DimUtilizador*.

Concluída a fase de extração dos dados, seguiu-se a etapa de limpeza e tratamento dos registos. Esta transformação assentou, essencialmente, nos passos esquematizados na **Figura 37**.

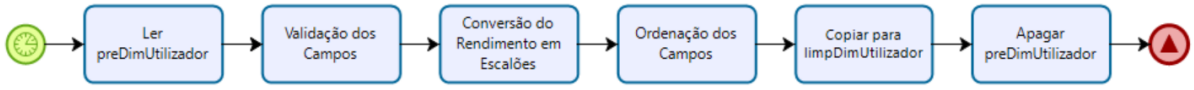


Figura 37 - Esquema BPMN da metodologia de transformação dos dados para o povoamento da *DimUtilizador*.

Numa primeira etapa, validaram-se os dados quanto ao seu tipo e valores possíveis a tomar, de acordo com as restrições impostas na **Tabela 16**.

Tabela 16 - Validação dos dados

Atributo	Tipo	Valores Possíveis
Sexo	*String*	{Feminino, Masculino}
Habilitações	*String*	{Básico, Secundário, Licenciatura, Mestrado, Doutoramento}
Rendimento (milhares)	Inteiro	[0, 500]
Operação	*String*	{Novo, Atualizado, Removido}

Posteriormente, o rendimento de cada utilizador foi agrupado em classes, de acordo com as seguintes correspondências:

- **Rendimento baixo** – valores de rendimento inferiores a 7500€ por ano.
- **Rendimento médio-baixo** – valores de rendimento entre 7500€ e 15000€ por ano.
- **Rendimento médio** – valores de rendimento entre 15000€ e 25000€ por ano.

- **Rendimento médio-alto** – valores de rendimento entre 25000€ e 50000€ por ano.
- **Rendimento elevado** – valores de rendimento superiores a 50000€ por ano.

Após o processamento indicado, os registos tratados foram copiados para a tabela *limpDimUtilizador* e a tabela *preDimUtilizador* foi apagada. De notar que no início do processo de limpeza, a descrição e a data e hora em que ocorreu a tarefa foram registadas numa tabela, que contém os relatórios de execução dos principais eventos do processo de ETL. Por sua vez, no final da transformação, também se efetuou o mesmo procedimento para se sinalizar o fim da tarefa. Além disso, no caso de existirem registos que falhem *steps* relacionados com a limpeza dos dados, esses registos são incorporados numa tabela de quarentena, com a respetiva justificação por terem sido eliminados. Este procedimento possibilita que, numa fase posterior ao processo de ETL estar concluído, o analista possa aceder e analisar o conteúdo das tabelas de quarentena e, com isso, decidir o destino final dos registos nelas contidos.

A implementação efetuada no Spoon, relativa a esta transformação, é a que se mostra na **Figura 38**. Esta transformação tem de ser, à semelhança da anterior, executada de forma transacional.

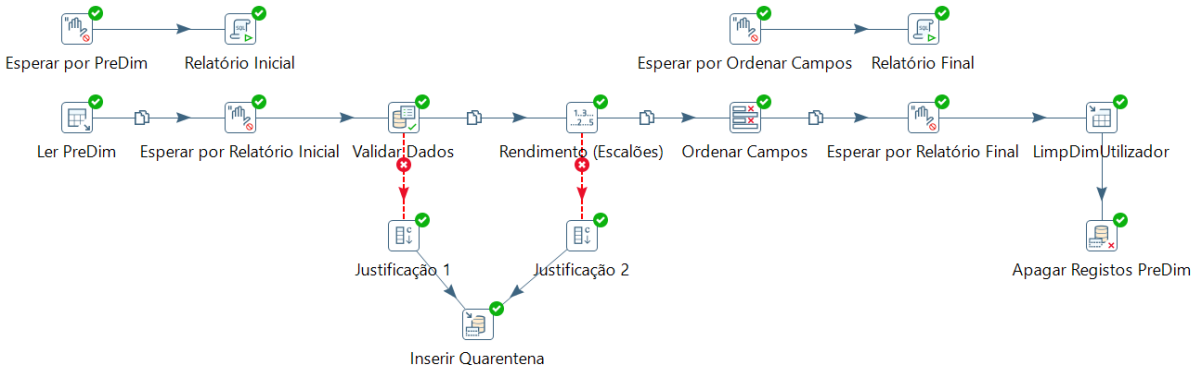


Figura 38 - Implementação, no Spoon, da transformação relativa ao processo de limpeza da *DimUtilizador*.

A *job* relativa ao carregamento da dimensão *Utilizador* relaciona-se com o processo de geração das *surrogate keys* (SK), para os novos registos, e com a integração e atualização dos registos tratados no DW. Assim, em primeiro lugar, foi preciso substituírem-se as chaves primárias dos utilizadores (*ids*) por chaves auto-incrementais inteiras, as SK, de forma a permitir-se estabelecer o relacionamento entre a tabela de dimensão e a tabela de factos. Ainda que, à primeira vista, não sejam evidentes os benefícios da utilização de SK, por oposição ao uso de chaves naturais, estas chaves apresentam vantagens consideráveis e devem ser sempre, salvo raras exceções, as chaves primárias preferíveis das tabelas de dimensão. As principais mais-valias da sua utilização prendem-se com o facto de não poderem ser reutilizáveis

(enquanto as chaves naturais podem ser recicladas) e de constituírem números inteiros relativamente pequenos, o que melhora significativamente o desempenho do sistema (Kimball and Caserta, 2004). Uma vez determinadas as SK para os novos registos, procedeu-se ao carregamento e à atualização das tabelas relativas às dimensões *Utilizador* e *UtilizadorHST* no DW. De uma forma genérica, a *job* responsável pelo carregamento da dimensão em estudo pressupõe a execução sequencial das três transformações representadas na **Figura 39**. Importa salientar o facto de cada uma destas transformações necessitar de ser concretizada de forma transacional.

Figura 39 - Constituição da *job Carregar Utilizador*.

As duas primeiras transformações dizem respeito à criação e integração das SK nos novos registos. Com o intuito de se evidenciarem as interligações entre estas duas transformações e elucidar acerca dos processos que ocorrem em cada uma delas, foi construído o esquema BPMN da **Figura 40**. De um modo sucinto, na transformação "Gerar SK", foi lida a última SK utilizada e definida uma variável, designada por *skey*, que inicialmente corresponde ao seu valor incrementado de uma unidade. Por sua vez, na transformação "SKGUtilizador", aos dados já limpos e transformados, foi acrescentada uma nova coluna referente à SK. Assim, para os registos novos, atribuíram-se SK geradas de forma sequencial, com início no valor da variável *skey*. Por outro lado, no caso dos registos atualizados ou removidos, como a sua SK correspondente já existe no DW, e não deve ser criada uma nova, associou-se a este parâmetro um valor omisso, do tipo *NULL*. No final de todos os registos terem sido processados, a maior SK atribuída aos novos registos foi guardada, de modo a possibilitar-se a repetição de todo este processo em futuros ciclos de ETL, tendo-se como ponto de partida este valor.

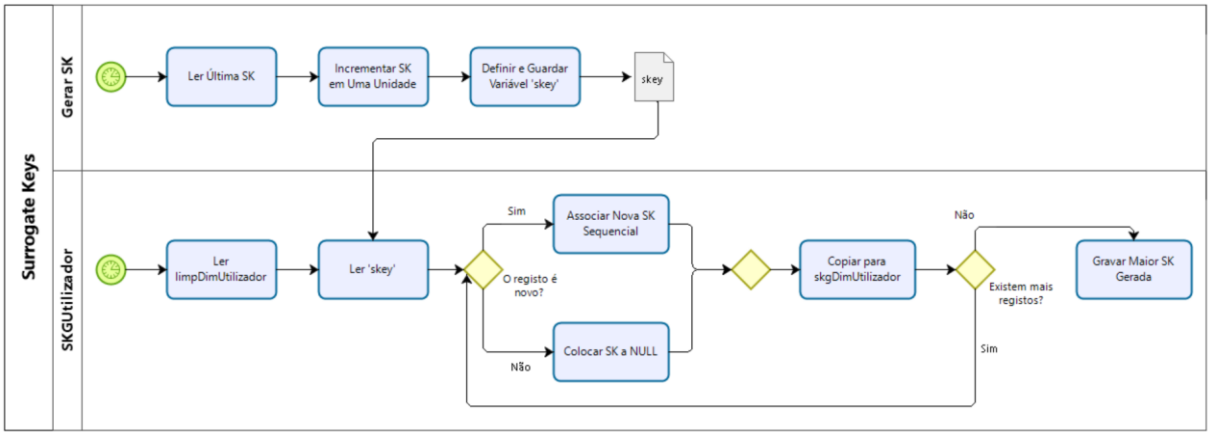


Figura 40 - Esquema BPMN do processo de criação das SK para o povoamento da *DimUtilizador*.

Uma vez geradas as SK, foi executada a transformação que se relaciona com o carregamento dos registos no DW. Como a dimensão *Utilizador* é propensa a variações pouco frequentes nos atributos relativos às habilitações literárias, rendimento e local de residência, esta dimensão é considerada uma *Slowly Changing Dimension* (SCD). Existem vários tipos de SCD, que se distinguem entre si consoante a estratégia adotada para lidar com a modificação dos valores dos atributos, sendo os mais comuns os tipos 1, 2, 3 e 4. De uma forma breve, as SCD do tipo 1, ao contrário das SCD dos restantes tipos, não preservam o histórico e os valores antigos são substituídos pelos atuais. Por sua vez, nas SCD do tipo 2, são acrescentadas três novas colunas relativas às datas de início e de expiração e ao estado do registo (ativo ou inativo). Assim, neste tipo de SCD, os registos antigos e os atuais partilham uma tabela comum e possuem SK diferentes. Já nas SCD do tipo 3, o valor antigo é substituído pelo novo, mas garante-se a conservação do histórico através da adição de uma nova coluna com o valor antigo. Na estratégia do tipo 4, é acrescentada uma tabela que guarda os registos históricos dos atributos que tenham sofrido alterações. Assim, neste caso, passam a existir duas tabelas, uma com os dados atuais e uma segunda tabela complementar, com os dados antigos. (Kimball and Ross, 2013) No caso da dimensão *Utilizador*, considerou-se importante armazenar o seu histórico de registos, de modo a que as consultas ao DW pudessem também ter em conta os dados antigos, que podem ser úteis para análises mais detalhadas ao comportamento dos índices dos utilizadores. Assim, foi preciso ponderar-se entre os tipos 2, 3 e 4, de modo a selecionar-se o mais indicado. Uma SCD do tipo 3 apresenta uma clara desvantagem em relação aos outros dois tipos, que se deve ao facto de apenas conseguir armazenar, no máximo, para um determinado atributo de um dado registo, o mesmo número de valores históricos que o de colunas adicionadas, sendo, por isso, limitada. Entre as SCD dos tipos 2 e 4, não há vantagem em utilizar-se uma dimensão do tipo 2, na medida em que a escolha deste tipo de SCD gera um acréscimo desnecessário de complexidade e implica a geração de novas chaves SK sempre que surgem registos modificados ou removidos. Deste modo, constatou-se que a melhor solução passaria por se considerar esta SCD como sendo do tipo 4, pelo facto de possibilitar o armazenamento dos dados históricos e apresentar um desempenho computacional otimizado, comparativamente aos restantes tipos. (Santos and Belo, 2011) Assim, para o carregamento no DW dos dados relativos aos utilizadores foi preciso atualizar e povoar-se, não só a tabela de dimensão, como também a correspondente tabela de histórico.

A etapa de carregamento dos registos no DW exigiu, neste caso, um planeamento acurado e organizado, capaz de assegurar um correto encadeamento entre os fluxos dos vários tipos de registos. É preciso salientar-se o facto de poderem existir vários registos associados a um

mesmo utilizador, uma vez que, durante o período de tempo entre o ciclo de ETL a ser executado e o ciclo anteriormente realizado, se permite que os utilizadores se registem, efetuem múltiplas atualizações aos seus dados e ainda cancelem os seus registos, sem que exista um limite às operações que podem exercer durante este intervalo. Deste modo, estas situações requerem um maior cuidado e uma ordem pré-definida do fluxo dos dados. Assim, numa primeira etapa, deve-se garantir que os dados do tipo "novo" são os primeiros a serem inseridos na dimensão *Utilizador*, com o campo relativo ao estado a "A" (ativo). Posteriormente, deve ser efetuado o tratamento dos dados atualizados, em duas fases distintas, de modo a que, inicialmente, sejam transferidos os dados antigos existentes na dimensão *Utilizador* para a tabela de histórico, e só depois de concluída esta etapa é que a tabela de dimensão deve ser atualizada com os novos dados. Para se inserirem os registos desatualizados na tabela de histórico é necessário ordenar-se os *ids* dos utilizadores do tipo "atualizado" de forma crescente e averiguar-se a existência de indivíduos que tenham efetuado múltiplas atualizações aos seus dados. No caso de se verificarem várias atualizações, é considerada apenas a mais recente (ou seja, o registo com uma maior data e hora de modificação), uma vez que, como a *DimUtilizador* se trata de uma dimensão com variação lenta, as atualizações anteriores não são relevantes e apenas a mais atual importa ser captada. Num passo seguinte, devem ser selecionados todos os *ids*, sem repetições, dos utilizadores que tenham procedido à atualização dos seus dados, para se possibilitar a sua identificação. Recolhidos estes *ids*, efetua-se a sua junção com a tabela de dimensão, pelo processo de igualdade entre chaves, com o intuito de se extraírem e transferirem os registos com os *ids* identificados para a tabela de histórico, passando estes registos a serem considerados desatualizados. A sua data de expiração é também assinalada como sendo a data/hora da modificação do registo mais recente, para esse *id*. Uma vez carregada a tabela de histórico é, então, possível proceder-se à atualização da *DimUtilizador*, a partir dos registos relativos à última atualização efetuada por cada um dos utilizadores. Para isso, é preciso indicar-se que estes registos devem substituir os registos existentes na tabela que tenham o mesmo *id*. Além disso, as atualizações apenas precisam de ser efetuadas ao nível dos parâmetros que apresentam possíveis variações, como é o caso das habilitações literárias, do rendimento e da localidade do utilizador, podendo manter-se as restantes variáveis com os valores de origem. O último tipo de dados a poder ser processado é o que se relaciona com os registos que foram removidos. Estes registos apenas podem ser marcados como inativos na tabela de dimensão após se ter concluído o processamento dos dados novos e dos atualizados. Para isso, associa-se um marcador "I" (inativo) a todos aqueles que sejam do tipo "removido" e atualiza-se a

DimUtilizador relativamente ao campo "estado", de forma a que os valores "A" destes registos sejam convertidos em "I".

A implementação efetuada no Spoon de todo o procedimento descrito, relativo à transformação "Carregar DW", está ilustrada na **Figura 41**. Desta forma, depois de executada esta transformação, a dimensão *Utilizador* e a sua correspondente tabela de histórico foram carregadas no DW.

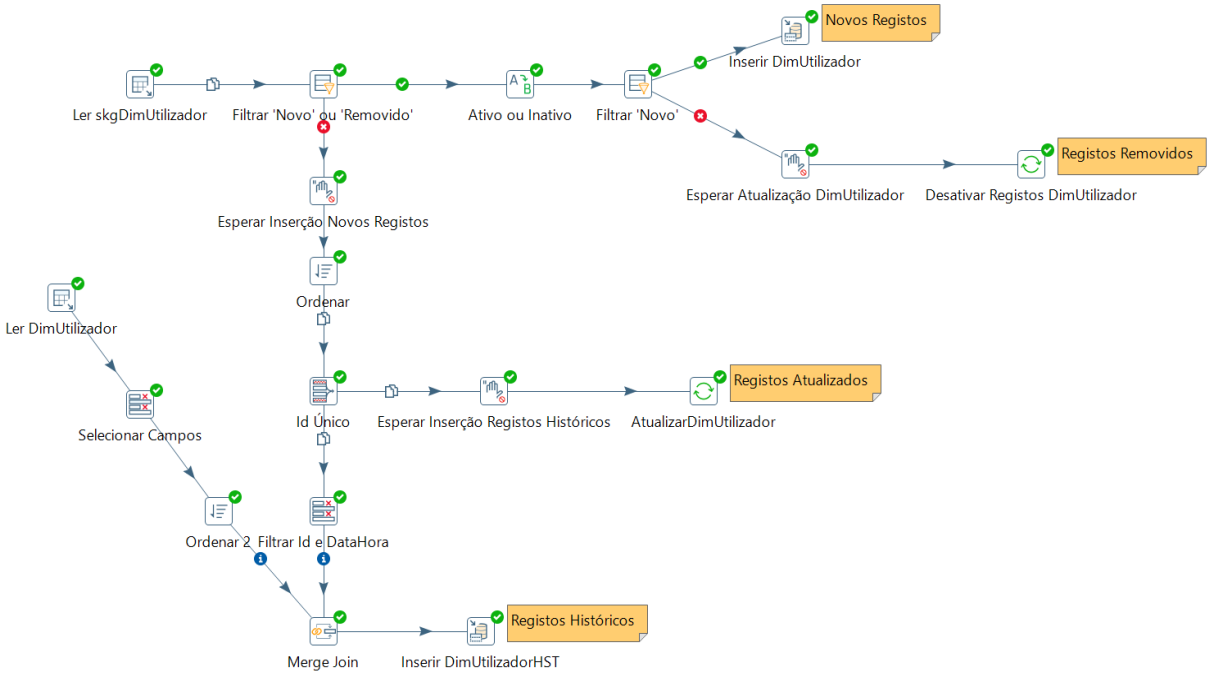


Figura 41 – Implementação, no Spoon, da transformação relativa ao processo de carregamento da *DimUtilizador*.

No que diz respeito ao povoamento das dimensões relativas ao calendário e aos distritos, o facto de elas dependerem de fontes de dados estáticas, simplificou consideravelmente o processo de ETL. Assim, como nestes casos, este processo apenas tem de ser realizado uma única vez, não foi necessário ter-se em atenção as situações de CDC, e o povoamento dos dados foi efetuado de forma total.

Analogamente ao processo de ETL da dimensão *Utilizador*, o da dimensão *Distrito* também requereu a criação de uma *job* de alto nível, similar à da **Figura 35**. No entanto, neste caso, esta *job* foi formada apenas pela transformação referente à extração dos dados e pela *sub-job* responsável pelo carregamento no DW. Assim, a transformação associada à limpeza foi dispensada, na medida em que os dados importados da fonte já se encontravam prontos a serem utilizados, sem necessidade de serem manipulados.

Em relação à extração dos dados da fonte, este procedimento foi efetuado de forma transacional e consistiu exclusivamente numa cópia integral dos seus registos para uma tabela na DSA, designada por *preDimDistrito*.

Por sua vez, o processo de carregamento, também ele transacional, teve um modo de funcionamento análogo ao da **Figura 39**. Tal como nesta figura, a primeira transformação lê e define a variável *skey*, que diz respeito à SK. Esta etapa prévia é necessária para que a segunda transformação (com um mecanismo idêntico ao da **Figura 40**), que se relaciona com o processo de atribuição de SK a cada registo, possa ser executada. Os passos criados no Spoon para a implementação desta transformação são os que se observam na **Figura 42**. No final, a última transformação tem como objetivo o carregamento dos dados resultantes na dimensão *Distrito*, através da migração dos dados da DSA para o DW.

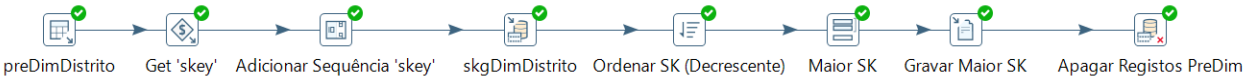


Figura 42 - Processo de *Surrogate Key Generator* relativo à *DimDistrito*.

Por outro lado, o povoamento da dimensão *Calendário* apresentou a particularidade de a fonte de dados poder ser gerada diretamente no Spoon e, por isso, o processo de manipulação e carregamento dos dados pôde ser concretizado numa única transformação. A primeira data do sistema foi definida como sendo o dia 02/01/2017, uma vez que correspondia à data do índice mais antiga contida na Fonte 2. Assim, além desta, foram ainda geradas mais datas, perfazendo-se um total de 7320, de modo a garantir-se uma validade do SDW de, aproximadamente, 20 anos. Tendo-se as datas criadas, tornou-se apenas necessário calcular a semana, mês, trimestre e ano correspondentes a cada uma delas, sendo que, no caso específico da semana, este valor foi convertido para o nome do dia da semana que lhe estava associado. Efetuado este procedimento, a dimensão *Calendário* foi, assim, povoada para o DW. Note-se que esta dimensão não necessitou da geração de SK, uma vez que, para estas tabelas, se recomenda o uso de *smart keys* e, consequentemente, a própria data pode ser usada como chave (Casters, Bouman and Dongen, 2010). Na **Figura 43** apresenta-se um esquema da metodologia utilizada para o povoamento desta dimensão.

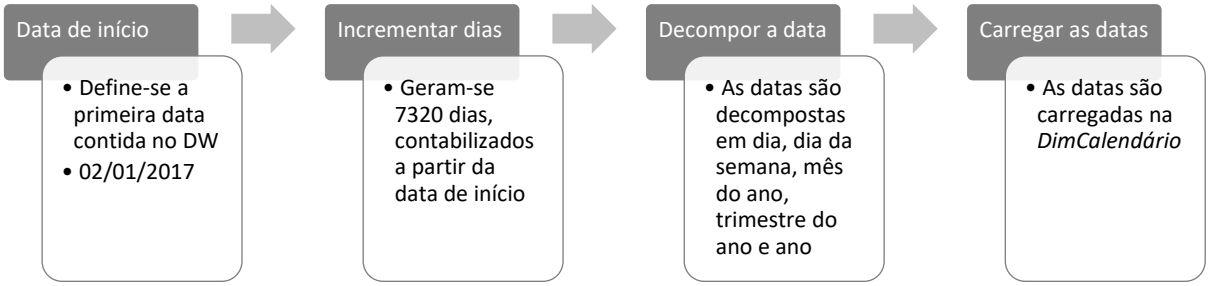


Figura 43 - Metodologia para o povoamento da *DimCalendário*.

4.5.4 ETL da Tabela de Factos

No que se refere à tabela de factos, apesar de esta tabela apenas poder ser carregada no DW depois de todas as dimensões terem sido inseridas, o seu processo de extração e de limpeza foi efetuado em paralelo com o carregamento das dimensões. Assim, o processo de ETL da tabela de factos foi decomposto em duas etapas assíncronas, relativas à execução das jobs *preTF* e *TFBemEstar*, representadas na **Figura 33** e na **Figura 34**, por esta ordem. Enquanto a primeira *job* não requer precedências, a segunda apenas pode ser efetuada após todas as outras tabelas terem sido carregadas e após a *job preTF* ter sido executada.

Relativamente à *preTF*, esta *job* divide-se ainda num conjunto de *sub-jobs* e de transformações transacionais, que dizem respeito aos processos de extração e de limpeza dos dados e à conciliação das fontes, tal como mostra a **Figura 44**.

Figura 44 - Constituição da *job preTF*.

Neste caso, como existem duas fontes de dados distintas para povoar esta tabela, o processo de extração dos dados foi também dissociado em duas etapas. Na primeira, extraíram-se os dados contidos na Fonte 2, que correspondem aos valores que variam diariamente e cujos registos se relacionam diretamente com uma linha da tabela de factos e, apenas posteriormente, puderam ser extraídos os dados necessários da Fonte 1, relativos a informações pessoais e clínicas, de menor variação, que também servem de base ao cálculo dos índices. Note-se que esta ordem foi definida de modo a que não fossem precisos extrair todos os registos da Fonte 1, mas apenas aqueles que estão intimamente relacionados com os registos dos indivíduos que têm índices para calcular (registos extraídos da Fonte 2), diminuindo-se, assim, significativamente o esforço computacional.

Para o processo de extração dos dados da Fonte 2, foi preciso, em primeiro lugar, definir-se a sua política de extração. Como nesta fonte de dados, os registos não podem ser modificados, uma vez que se tratam de transações diárias, que apenas podem ser acrescentadas em novas linhas da fonte, considerou-se que a metodologia de extração mais apropriada para esta situação seria a de se retirarem os registos com uma data superior à da última instância extraída. Este processo pode ser executado de uma forma não intrusiva, na medida em que a própria fonte de dados já dispõe das datas em que os registos aí foram incluídos. Assim, para a identificação da data do último registo extraído, foi criada, através do Spoon, uma tabela de *log*

na DSA, que contém, entre outros parâmetros, a data de início (*STARTDATE*) e a data de fim (*ENDDATE*). Na **Tabela 17** está representado um extrato da tabela de *log* gerada. A data de início corresponde à data em que o último registo anteriormente extraído tinha sido colocado na Fonte 2, enquanto a data de fim identifica a data de inserção do último registo contido, atualmente, nesta fonte de dados. Desta forma, sempre que seja ativado um novo processo de ETL, a data de fim do último processo passa a corresponder à nova data de início. Importa notar que, no caso da primeira extração, a data de início é automaticamente associada pelo sistema como sendo o dia 31/12/1899 e, deste modo, o período de início não se torna uma limitação.

Tabela 17 - Extrato da tabela de *log* que é criada na DSA

ID_BATCH	CHANNEL_ID	TRANSNSAME	STATUS	…	STARTDATE	ENDDATE	…
1	7b852dd0-6636-4fce-8cd4-efd8ae7b9aa0	ExtrairTF2	end	…	1899-12-31 23:00:00.000	2019-09-03 00:00:00.000	…

A implementação prática desta estratégia de extração de dados foi efetuada no Spoon, de acordo com os *steps* que se visualizam na **Figura 45**. Além disso, para a extração exclusiva dos registos que ainda não foram incorporados na tabela de factos do DW foi declarada a condição *WHERE* que se mostra na figura. Neste contexto, os pontos de interrogação representam as variáveis *STARTDATE* e *ENDDATE*, por esta ordem, que vão sendo atualizadas sempre que um novo processo de ETL for desencadeado. Deste modo, a política de extração adotada tanto é válida em situações de povoamento total como de povoamento incremental. Uma vez extraídos, os dados foram armazenados numa tabela da DSA, designada por *preTF2*.

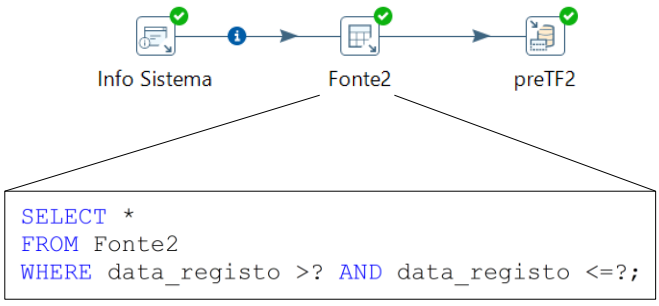


Figura 45 - Implementação, no Spoon, da transformação relativa ao processo de extração dos dados da Fonte 2 da *TFBemEstar*.

Uma vez concluído o processo de extração dos novos registos da Fonte 2, foram retiradas da Fonte 1 apenas as instâncias que diziam respeito aos utilizadores que tinham novos registos na Fonte 2. Estes dados da Fonte 1 têm apenas como finalidade fornecer parte das métricas que suportam o cálculo do índice de bem-estar. Importa notar que, neste caso, não se optou por se reproduzir integralmente os dados da Fonte 1 para uma tabela acessória na DSA, que pudesse fornecer os registos necessários quando solicitado, pelo facto de os registos da Fonte 1 serem sujeitos a atualizações periódicas nos campos relativos às análises médicas. Desta forma, como a estratégia adotada recolhe os dados diretamente da fonte, os seus valores já são, garantidamente, os atuais e, por isso, as modificações nestes dados deixam de ser um problema. Relativamente à forma de se extraírem os dados, a estrutura do processo implementado foi semelhante à da **Figura 45**, tendo-se agora como *input* da Fonte 1 os *ids* ordenados, sem repetições, da Fonte 2. Para a extração dos dados da Fonte 1, para a tabela *preTF1*, a condição em SQL acrescentada foi "*WHERE id = ?*".

A fase seguinte consistiu na execução da *job* responsável pela limpeza dos dados de ambas as fontes. Ao contrário da *job* anterior, relativa à extração dos dados, esta *job* permite que as transformações associadas aos dados das fontes 1 e 2 possam ser executadas em paralelo, sem a necessidade de se definir uma ordem específica de concretização.

A transformação dos dados extraídos da Fonte 1 consistiu na execução dos seguintes passos principais:

1. Leitura dos registos da *preTF1*.
2. Apuramento do número de registos existentes para o mesmo *id*. Os casos em que os mesmos *ids* aparecem em vários registos são encaminhados para a tabela *QuarentenaTF*, com a justificação "Registo com duplicação de dados".
3. Verificação do tipo de dados e validação dos seus valores. Os registos que falhem este *step* são inseridos na tabela *QuarentenaTF*, com a justificação "Dados inconsistentes".
4. Cálculo dos campos relativos à idade, IMC e quantidade de cigarros fumada ao longo da vida, de acordo com as expressões:
    - Idade: **DATEDIF** ([Data Nascimento]; **today**(); **"y"**)
    - IMC: [Peso] / ([Altura] / **100**) ^ **2**
    - Quantidade Cigarros: [Num Cigarros] * [Num Anos Fum]* **365** / **1000**

5. Teste à coerência dos campos relativos ao tabagismo, através da utilização das condições indicadas na **Figura 8**. Em caso de falha, os registos são alocados na tabela de quarentena, associados à justificação "Incoerência de dados".
6. Inserção dos registos finais na tabela *limpTF1*.
7. Remoção de todo o conteúdo da tabela *preTF1*.

Por sua vez, a transformação dos dados extraídos da Fonte 2 implicou a adoção da seguinte metodologia:

1. Leitura dos registos da *preTF2*.
2. Filtragem apenas do registo mais recente no caso de um mesmo utilizador ter, para a mesma data, múltiplos registos, sendo os restantes ignorados.
3. Verificação do tipo de dados e validação dos seus valores. Os registos que falhem este *step* são inseridos na tabela *QuarentenaTF*, com a justificação "Dados inconsistentes".
4. Confirmação de que a data do registo na fonte de dados é maior, ou igual, que a data da recolha dos parâmetros relativos às pressões arteriais e calorias (data do índice). Para tal, é preciso verificar-se a condição "*data_registo* > = *data_índice*". No caso de existirem registos que não cumpram esta restrição, são encaminhados para a tabela *QuarentenaTF* para mais tarde serem examinados pelo analista.
5. Conversão das calorias em escalões relacionados com o nível de exercício físico praticado. Para isso, considerou-se que "nenhum" exercício corresponderia a menos de 200 calorias despendidas, "baixo" a valores entre 200 e 500, "moderado" ao intervalo entre 500 e 800, e, por último, "elevado", a mais de 800 calorias, de acordo com a fórmula dada por:
    - **IF** ([calorias] < **200**; **"Nenhum"**; **IF** ([calorias] < **500**; **"Baixo"**; **IF** ([calorias] < **800**; **"Moderado"**; **"Elevado"**)))
6. Inserção dos registos finais na tabela *limpTF2*.
7. Remoção de todo o conteúdo da tabela *preTF2*.

Depois de extraídos e transformados os dados de ambas as fontes, foi preciso efetuar a sua conciliação. Este processo iniciou-se pela junção destes registos através da igualdade entre os seus *ids*, utilizando-se, para isso, um *RIGHT OUTER JOIN*. Este tipo de junção permitiu manter todo o universo de registos da tabela *limpTF2* e conciliá-lo com a *limpTF1*. Os principais passos de tratamento efetuados nesta fase prenderam-se com o encaminhamento para a tabela de quarentena das transações provenientes da Fonte 2 relativas a utilizadores não

registados na Fonte 1 e dos registos pertencentes a utilizadores que não tenham atualizado as suas análises médicas há mais de um ano, ou que, em contrapartida, tenham registos de análises mais recentes que o do índice a calcular. Com isto, assegurou-se que os índices dizem respeito apenas a utilizadores registados e, além disso, garantiu-se que os dados relativos às análises correspondem a valores de medições efetuadas com uma diferença menor do que um ano, comparativamente com a data da avaliação do valor do índice e, por isso, são suficientemente recentes para poderem ser considerados no seu processo de cálculo. Por outro lado, também se impediu que a data das análises fosse superior à do índice. Com vista a filtrarem-se os registos de acordo com este intervalo máximo de um ano entre a data das análises e a data dos índices e a garantir-se uma data das análises superior à do índice, recorreu-se às condições descritas, em Java, na **Figura 46**. No final desta transformação, os dados foram carregados numa tabela designada por *concTF* e o conteúdo das tabelas *limpTF1* e *limpTF2* foi apagado.

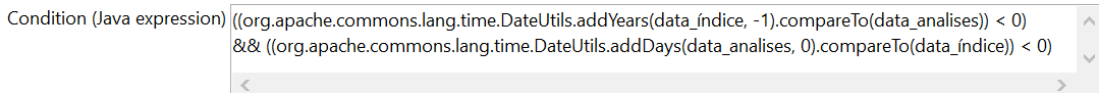


Figura 46 - Condições, em Java, para garantir a validade máxima de um ano entre as datas do índice e das análises e que a data do índice é superior à das análises.

Com o término do processo de conciliação das fontes de dados relativo ao povoamento da tabela de factos, concluiu-se, igualmente, a execução da *job preTF*. Tendo-se também efetuado o carregamento das tabelas de dimensão no DW, encerrou-se a primeira parte da *job ETL* e, por conseguinte, a *job TFBemEstar*, assinalada na **Figura 34**, pôde ser iniciada. A estrutura mais genérica desta *job* é a que se representa na **Figura 47** e pressupõe uma execução transacional das suas transformações. Mais uma vez, realça-se o facto de o carregamento prévio das dimensões ser essencial para o correto povoamento da tabela de factos, na medida em que, só após estas tabelas estarem devidamente inseridas no DW, é que se pode proceder à conciliação da tabela de factos com as dimensões, através de chaves comuns.

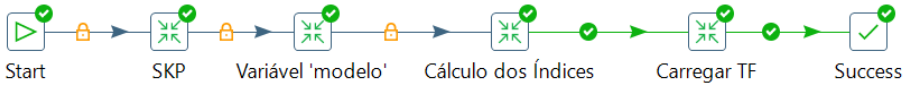


Figura 47 - Constituição da *job TFBemEstar*.

A segunda etapa da job *ETL* iniciou-se pelo processo de *Surrogate Key Pipeline*, que consistiu numa sequência de operações do tipo *lookup*, no sentido de se identificarem as SK relativas às dimensões *Utilizador* e *Distrito*, correspondentes a cada registo. No caso de não existir uma correspondência entre a chave natural indicada numa transação e uma SK, esse registo é encaminhado para a tabela de quarentena. Relembre-se que, no caso da dimensão

Calendário, não foram geradas SK e usou-se a própria data como *smart key*. Assim, em consequência disso, a tarefa de *lookup* não se aplicou a esta dimensão. Os passos efetuados para a implementação desta transformação no Spoon podem ser observados na **Figura 48**.

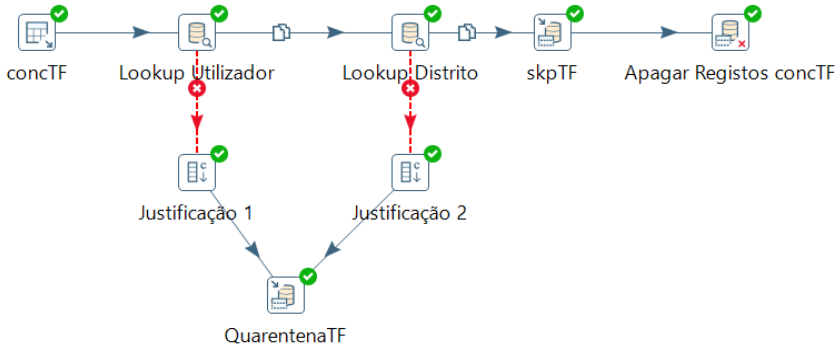


Figura 48 - Processo de *Surrogate Key Pipeline*.

Depois de efetuada a correspondência das chaves naturais com as SK, foi criada uma nova variável, designada por "modelo", representativa do nome do melhor algoritmo de DM encontrado para o cálculo dos índices. Esta variável foi lida com base no ficheiro resultante do processo de determinação dos *scores* globais dos modelos de DM construídos (**Figura 14**), que contém o nome da técnica com a pontuação mais elevada. Deste modo, tendo-se selecionada a técnica de DM a aplicar, foi calculado, numa primeira fase, o valor probabilístico (entre 0 e 1) de cada transação, no que diz respeito ao desenvolvimento de DCV. Para tal, para este cálculo, foi preciso indicar-se o modelo de DM a ser importado e, para isso, inseriu-se o *path* do ficheiro como sendo do tipo **diretoria*/${modelo}_1_1.model*. Desta forma, como os ficheiros dos algoritmos de DM criados são designados pelo seu próprio nome seguidos da expressão "_1_1", o sistema automaticamente utiliza aquele que corresponde à maior pontuação global de acurácia e de sensibilidade. Posteriormente, de forma a converter-se as probabilidades obtidas em índices, foi utilizada a fórmula dada pela **Equação 6**.

Determinados os índices, o último passo consistiu na migração dos dados da DSA para o DW, de forma a carregar-se a tabela de factos e a concluir-se, com sucesso, o processo de ETL.

4.5.5 Validação e Considerações Finais do SDW Implementado

Com o intuito de se testar o SDW desenvolvido foi efetuada, em primeiro lugar, uma revisão genérica de todas as etapas e, posteriormente, foi executado o primeiro povoamento ao sistema, considerando-se os dados da fonte relativa ao DM para alimentar a *job DM* e os dados das fontes 1 e 2, necessários para a execução da *job ETL*.

Em relação às *jobs* principais associadas aos processos de DM e de ETL, após estas terem sido desencadeadas e terminadas, foi previsto o envio de mensagens de *email* informativas

acerca do estado de sucesso da sua conclusão. No caso da *job DM*, o "sucesso" ocorre quando um modelo de DM é criado e selecionado, e a sua designação é gravada num ficheiro de texto. Por outro lado, no caso de insucesso, o *email* enviado indica o motivo pelo qual não sucedeu uma conclusão efetiva da *job*, que se pode dever a uma de duas situações:

1. Não existir, pelo menos, 5000 registos na fonte relativa ao DM.
2. Ocorrência de erro numa das etapas de execução do DM.

Após a execução desta *job*, foi recebido um *email* de sucesso e, por isso, pôde-se validar o seu funcionamento. A título de exemplo, no Anexo IX, mostra-se um excerto do *email* enviado para o analista em caso de sucesso da *job* em estudo.

Relativamente à conclusão do processo de ETL, a sua execução também é testada pela receção de um *email* a informar o "sucesso". Neste caso, considera-se "sucesso" sempre que o ETL tenha sido devidamente concluído e o povoamento do DW tenha ocorrido de forma efetiva em todas as suas estruturas dimensionais. Quando isto acontece, é enviado um *email* de confirmação, tal como mostrado no Anexo IX, em que se indicam a data de conclusão do processo e os tempos detalhados de início e fim associados à execução de cada *job* e transformação. De um modo análogo, sempre que o procedimento for considerado como "insucesso", é enviado um *email* com uma descrição do erro e da etapa em que ocorreu. Depois de executada esta *job*, considerando-se o povoamento total das fontes 1 e 2, o *email* recebido validou o sucesso da sua conclusão.

No que diz respeito ao povoamento das tabelas do DW, verificou-se que todas elas estavam devidamente preenchidas. A título de exemplo, na **Figura 49** pode-se observar a dimensão relativa ao utilizador após este primeiro povoamento, em que se verifica que todos os campos foram corretamente inseridos, incluindo-se o das SK geradas. Além disso, também se confirmou que, neste carregamento, que apenas diz respeito a novos registos, todas as instâncias apresentam o estado a "A".

	sk	id_utilizador	sexo	habilitacoes	rendimento	local	estado
1	4469	1	Masculino	Licenciatura	Médio-Alto	PRT	A
2	4470	2	Masculino	Mestrado	Baixo	LSB	A
3	4471	3	Feminino	Básico	Médio	PRT	A
4	4472	4	Masculino	Licenciatura	Baixo	AVR	A
5	4473	5	Masculino	Secundário	Médio	GRD	A
6	4474	6	Feminino	Básico	Médio-Alto	PRT	A

Figura 49 - Excerto da dimensão *Utilizador* após o primeiro povoamento.

Além desta dimensão, também se constatou que a tabela de factos e as dimensões referentes ao calendário e ao distrito apresentaram registos com valores dentro do expectável.

No caso da tabela de factos, observou-se ainda que o valor do índice relativo a cada registo tinha sido adequadamente inserido, variando entre -5 e 5.

Quanto às tabelas da DSA, verificou-se que todas elas estavam vazias e preparadas para a execução de um novo ciclo de ETL, com exceção da tabela de *log*, que continha o registo ilustrado na **Figura 50**, que marca a última data relativa à extração da Fonte 2, e com exceção das tabelas de quarentena e da tabela de auditoria, que poderia conter registos que entretanto tivessem sido inseridos na Fonte 1.

	ID_BATCH	CHANNEL...	TRANSN...	STATUS	LINE...	LINE...	LINES_U...	LINES_INPUT	LINES...	LINES...	ERRORS	STARTDATE	ENDDATE
1	1	90487d88-...	ExtractT...	end	0	0	0	100000	0	0	0	1899-12-31 ...	2019-10-15...

Figura 50 - Excerto da tabela de *log* relativa às tarefas de extração da Fonte 2.

No que toca às tabelas de quarentena, verificou-se que estas tabelas referentes ao utilizador e à tabela de factos incorporaram um conjunto significativo de registos que não satisfizeram as restrições impostas, com o campo relativo ao motivo devidamente indicado, para facilitar a visualização por parte do analista. Como exemplo da tabela de quarentena resultante do povoamento da tabela de factos apresenta-se um excerto da sua constituição na **Figura 51**.

hipotiroidismo	id_utilizador	calorias	pa_alta	pa_baixa	data_índice	data_registo	motivo	idade	IMC	exercício_físico
Não	NULL	NULL	NULL	NULL	NULL	NULL	Incoerência de Dados	34	31,7739318550568	NULL
NULL	NULL	NULL	135	95	2018-09-25 00:00:00.000	NULL	Utilizador não Registado	NULL	NULL	Elevado
Não	NULL	NULL	122	85	NULL	NULL	Data Análises Expirada ou Data Análises > Data Í...	57	21,3452728029988	Nenhum
Sim	NULL	NULL	143	100	NULL	NULL	Data Análises Expirada ou Data Análises > Data Í...	53	29,2421086364365	Elevado
Não	NULL	NULL	106	79	NULL	NULL	Data Análises Expirada ou Data Análises > Data Í...	54	36,6670209374004	Moderado

Figura 51 - Excerto da tabela de quarentena relativa à tabela de factos.

Em relação aos tempos de execução, através dos *emails* rececionados, foi possível constatar que a *job DM* apresentou um tempo total de execução de 5 s, enquanto a *job ETL* teve uma duração total de 8 min 45 s. Para a janela de oportunidade, definiu-se um tempo de 1h, de forma a permitir-se que nesse período de tempo se pudesse executar integralmente todo o processo de ETL, incluindo-se possíveis falhas que pudessem ocorrer. Além disso, considerou-se um ciclo de execução 24/24.

De forma a validar-se também o sistema durante a execução de um processo de ETL incremental, foram executadas as seguintes ações de modificação à Fonte 1, de acordo com a seguinte ordem:

1. Inserção de um novo registo, com *id* = 5043.
2. Alteração do local do registo com *id* = 4, passando de “AVR” para “PRT”.
3. Alteração das habilitações literárias e do distrito do registo com *id* = 4, passando-os para “Mestrado” e “LSB”, respetivamente.

4. Alteração do rendimento com *id* = 2, passando-o de "7" para "14".
5. Remoção do registo com *id* = 2.
6. Remoção do registo com *id* = 5.

Após a execução deste conjunto de etapas, a tabela de auditoria detetou as alterações efetuadas à Fonte 1, indicando o tipo de operação realizada e a data e hora da modificação, tal como se observa na **Figura 52**.

	id_operacao	id_utilizador	sexo	habilitacoes	rendimento	local	operacao	datahora
1	22052	5043	Feminino	Mestrado	10	BRG	novo	2019-10-17 15:07:50.573
2	22053	4	Masculino	Licenciatura	4	PRT	atualizado	2019-10-17 15:12:46.320
3	22054	4	Masculino	Mestrado	4	LSB	atualizado	2019-10-17 15:13:37.510
4	22055	2	Masculino	Mestrado	14	LSB	atualizado	2019-10-17 15:14:01.880
5	22056	2	Masculino	Mestrado	14	LSB	removido	2019-10-17 15:15:17.273
6	22057	5	Masculino	Secundário	24	GRD	removido	2019-10-17 15:15:26.720

Figura 52 - Conteúdo da tabela de auditoria após um conjunto de modificações à Fonte 1.

Deste modo, quando se desencadeou um novo processo de ETL, foram extraídos os registos da tabela de auditoria e as tabelas associadas à dimensão *Utilizador* e ao seu histórico foram atualizadas de acordo com a **Figura 53** e a **Figura 54**, respetivamente.

Da tabela da **Figura 53** verificou-se que o estado dos registos removidos passou a "I" e que, nos casos em que existiam múltiplas operações de atualização para o mesmo utilizador, como era o caso do utilizador com *id* = 4, apenas as mais recentes foram incorporadas.

	sk	id_utilizador	sexo	habilitacoes	rendimento	local	estado
1	4469	1	Masculino	Licenciatura	Médio-Alto	PRT	A
2	4470	2	Masculino	Mestrado	Médio-Baixo	LSB	I
3	4471	3	Feminino	Básico	Médio	PRT	A
4	4472	4	Masculino	Mestrado	Baixo	LSB	A
5	4473	5	Masculino	Secundário	Médio	GRD	I
6	4474	6	Feminino	Básico	Médio-Alto	PRT	A

Figura 53 - Excerto da dimensão *Utilizador* após o povoamento incremental.

Por outro lado, da **Figura 54** constatou-se que o registo antigo relativo ao utilizador com *id* = 4 foi inserido na tabela de histórico, com a data de referência igual à data/hora em que tinha sido colocado na tabela de auditoria. Além disso, em relação ao utilizador de *id* = 2, verificou-se que, apesar de ter existido uma operação de atualização associada a este utilizador, não foi colocado qualquer registo seu no histórico, uma vez que a última operação tinha sido a de remoção do registo.

	registo	sk	id_utilizador	habilitacoes	rendimento	local	dataref
1	1	4472	4	Licenciatura	Baixo	AVR	2019-10-17 15:13:37.510

Figura 54 - Histórico do utilizador após o povoamento incremental.

Em relação à Fonte 2 verificou-se também que, mediante um cenário de povoamento incremental, apenas os novos registos desta fonte eram extraídos. Deste modo, foi inserida uma nova linha na tabela de *log*, com o *STARTDATE* e o *ENDDATE* atualizados.

Uma vez realizadas todas estas etapas, o SDW implementado pôde ser validado e, nesse sentido, pôde-se proceder à construção de um cubo OLAP, para permitir um acesso mais simplificado aos dados. Com a conceção deste cubo foi, assim, iniciada a etapa final do sistema implementado, com a transformação dos dados em conhecimento. O resultado de todo o processo descrito, desde as fontes de dados até à criação dos cubos OLAP, nada mais é do que um pequeno passo para os dados. Por sua vez, os *dashboards* gerados a partir deles serão um grande passo para o conhecimento.

5. UMA VISUALIZAÇÃO INTERATIVA DOS RESULTADOS

Os cubos OLAP funcionam como uma forma eficiente de se organizarem estruturas multidimensionais e de se disponibilizarem os dados de um modo mais rápido, de acordo com o tipo de consultas mais frequentes. No caso da presente dissertação, um cubo deste tipo permite servir de base para a construção de gráficos intuitivos, que satisfazem os requisitos de exploração enumerados no capítulo anterior, ao responder a questões do tipo:

- Qual é o valor do índice de bem-estar cardíaco do utilizador X, no dia Y?
- Como é que o índice do utilizador X evolui ao longo do tempo?
- Qual é o índice global dos utilizadores?
- Como é que o índice global dos utilizadores evolui ao longo do tempo?
- Qual é o índice global dos indivíduos do sexo masculino? E do sexo feminino?
- Quais são os distritos do país que registam os piores e os melhores índices?

5.1 A Construção do Cubo OLAP

Para a conceção do cubo OLAP, em primeiro lugar, criou-se um novo projeto do módulo *Analysis Services* no Visual Studio, sintonizando-o ao DW do SQL Server. Posteriormente, depois de definidas a fonte de dados e as vistas da fonte de dados, procedeu-se à implementação do cubo, selecionando-se, para isso, as tabelas de dimensão e a tabela de factos do DW. Numa etapa seguinte, foi preciso especificarem-se ainda os atributos e as hierarquias de cada dimensão e as métricas a serem consideradas. Em relação às métricas, selecionaram-se apenas as medidas necessárias da tabela de factos para o suporte à determinação dos índices, ou seja, o índice não ponderado agregado como uma média ao longo do tempo, o índice não ponderado agregado como uma soma e o número de linhas distintas.

O cubo criado tem, essencialmente, de servir de base à construção de gráficos e de elementos visuais relativos ao índice por utilizador e aos índices gerais. Estes dois tipos de índices precisam de ser determinados de um modo específico para cada caso.

Particularizando-se para os índices relativos aos utilizadores individuais, estes indicadores devem refletir um valor ponderado ao longo do tempo, uma vez que é possível que, em dias ocasionais, alguns valores dos parâmetros de avaliação (como as pressões arteriais) apresentem grandes oscilações e, com isso, valores atípicos. Deste modo, os valores extremos devem ser atenuados e o histórico dos valores anteriores dos índices deve ser tido em conta para a determinação do valor atual. Assim, considerou-se que o valor que melhor representa o estado

da saúde cardiovascular de um dado indivíduo é o seu último valor de índice ponderado, que reflete, para cada instante, todos os valores dos índices anteriores. Para se poder disponibilizar este índice nos *dashboards*, foi preciso criar-se, no cubo, um membro calculado, que funciona como uma nova medida. Para isso, recorreu-se à instrução MDX indicada na **Figura 55**, que calcula a reta de regressão linear e devolve o valor de Y (índice ponderado) para um determinado X (data). Note-se que, neste caso, a medida "Índice" utilizada corresponde ao índice agregado pela média ao longo do tempo.

```
LinRegPoint(
Rank([Dim Calendário].[HierarquiaAno].CurrentMember, [Dim
Calendário].[HierarquiaAno].CurrentMember.Level.MEMBERS),
{[Dim Calendário].[HierarquiaAno].CurrentMember.Level.Members}.Item(0):[Dim
Calendário].[HierarquiaAno].CurrentMember,
[Measures].[Índice],
Rank([Dim Calendário].[HierarquiaAno].CurrentMember, [Dim
Calendário].[HierarquiaAno].CurrentMember.Level.MEMBERS)
)
```

Figura 55 - Instrução, em linguagem MDX, para a determinação do membro calculado *LinRegPoint*.

Este novo membro apresenta algumas limitações, na medida em que, sendo calculado através de uma regressão linear, o primeiro valor do índice devolve um resultado impossível de determinar e, além disso, também é possível que o valor retornado exceda a escala considerada de -5 a 5. Para se ultrapassarem estas fraquezas, foi então especificado o membro calculado do índice ponderado, dado pela **Figura 56**.

```
IIF([Measures].[LinRegPoint]>5, 5 ,
IIF([Measures].[LinRegPoint]<-5, -5 ,
    IIF([Measures].[LinRegPoint]<5 and [Measures].[LinRegPoint]>-5,
    [Measures].[LinRegPoint],
        [Measures].[Índice])))
```

Figura 56 - Instrução, em linguagem MDX, para a determinação do membro calculado *ÍndicePonderado*.

Por sua vez, para a determinação dos índices gerais, considerou-se que o valor mais adequado era um índice global, construído a partir do valor médio dos índices dos utilizadores e, para isso, foi criado o novo membro calculado, designado por *ÍndiceGlobal*, tal como se apresenta na **Figura 57**. Importa notar que, nesta situação, o índice usado foi o que se agrega pela soma.

```
[Measures].[Índice - TF]/[Measures].[TF Count]
```

Figura 57 - Instrução, em linguagem MDX, para a determinação do membro calculado *ÍndiceGlobal*.

Depois de o cubo estar devidamente construído, foi possível importá-lo, em tempo real, para o Power BI, para se poder proceder à visualização gráfica dos resultados. Esta visualização marca o início da apreensão do conhecimento, que ocorre quando a informação é transmitida e entendida por cada um dos utilizadores. Para a obtenção de informação útil a partir dos índices

calculados, foram gerados *dashboards*, cujos resultados foram organizados em três páginas distintas, uma direcionada para o utilizador individual e para profissionais de saúde, outra para fins estatísticos e uma última página para se possibilitar a visualização das tabelas de histórico.

5.2 *Dashboards* para Utilizadores Individuais e Profissionais de Saúde

Para a página destinada à consulta por parte de utilizadores individuais e de profissionais de saúde, foi preciso incluir-se um menu para se poder selecionar o *id* do utilizador que se pretende consultar. Para isso, foi aplicado um filtro à página de forma a que os *ids* propostos fossem apenas relativos a utilizadores com estado ativo, descartando-se, assim, todos os utilizadores com um registo cancelado.

A nível visual, elaborou-se um gráfico temporal, com a indicação do valor do índice ponderado ao longo do tempo para o utilizador em análise, em que cada valor corresponde ao valor do índice ponderado para essa data, tendo apenas em conta os registos anteriores. Além disso, incluiu-se também um indicador com o valor do índice ponderado associado à cor a que pertence (verde, amarelo, laranja ou vermelho). Para se associar de forma automática as cores aos valores do índice, recorreu-se a um mecanismo de formatação condicional, através da definição de um conjunto de regras aplicado ao nível da cor da letra e dos ícones.

Como exemplo, selecionando-se o utilizador cujo *id* é 2961, obtêm-se os resultados apresentados na **Figura 58**, considerando-se a data mais recente.

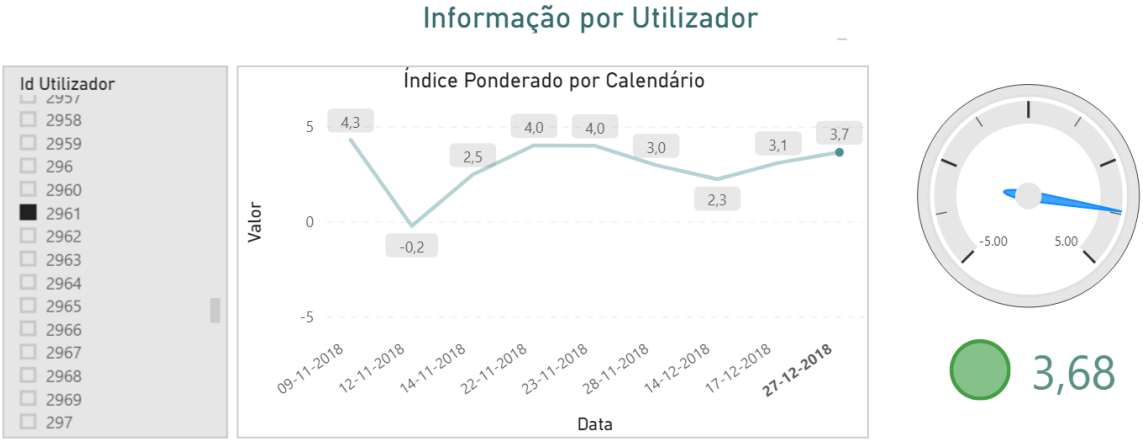


Figura 58 - *Dashboards* e indicadores relativos ao registo mais recente do utilizador de *id* 2961.

Do gráfico temporal da figura pode-se observar que este utilizador em concreto tem obtido, ao longo do tempo, bons valores relativos ao seu índice ponderado, com exceção do seu segundo registo, que teve um resultado negativo. Quando existem poucos registos, podem detetar-se mais facilmente mudanças acentuadas de valores dos índices ponderados mas, à

medida que mais registos são adicionados, mais estável se torna o sistema e menos acentuadas são as variações entre índices contíguos. Desta forma, os primeiros valores dos índices podem não ser credíveis e só depois de um período significativo de medições é que os valores começam a ser dotados de uma maior confiabilidade.

Da **Figura 58** retira-se ainda que o índice ponderado atual do utilizador em estudo é elevado (3.68), associando-se a uma cor verde. No entanto, o posicionamento do mostrador permite visualizar que ainda é possível melhorar este valor, servindo, assim, como uma motivação adicional para que o utilizador tente, ainda mais, aperfeiçoar os seus resultados.

Uma das maiores vantagens da criação de *dashboards* deste tipo é a grande interação que proporciona aos seus utilizadores. Caso, por exemplo, o utilizador pretenda saber o seu índice de uma forma anual, o próprio gráfico permite uma navegação intuitiva na hierarquia, sem que seja necessário criar outro gráfico para esse efeito. Além disso, se, por exemplo, o utilizador clicar sobre uma data em específico, o mostrador é atualizado de um modo instantâneo e devolve o valor e a cor associada à data selecionada. A título exemplificativo, caso o utilizador de *id* 2961 pretenda conhecer mais detalhes acerca do seu índice no dia 14 de dezembro de 2018 e clique nessa data, os resultados passam para o que se mostra na **Figura 59**.

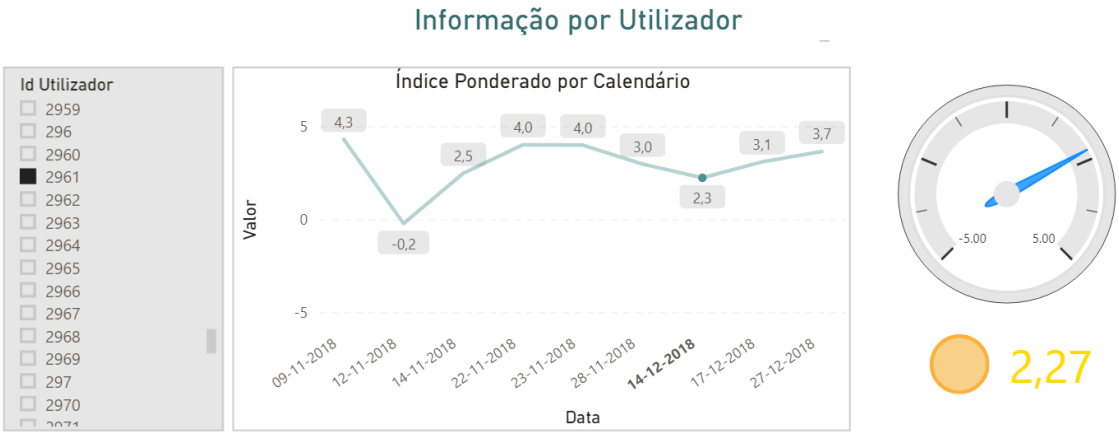


Figura 59 - *Dashboards* e indicadores relativos ao registo do dia 14/12/2018 do utilizador de *id* 2961.

Da visualização dos resultados, pode-se constatar que o índice obtido nesse dia mudou para a região amarela e que, por isso, em média, era pior do que os restantes valores.

5.3 *Dashboards* para Fins Estatísticos

Além dos *dashboards* relativos à consulta da evolução do índice dos utilizadores individuais, foram construídos gráficos, numa página distinta, vocacionados para os decisores que intervêm em matérias relacionadas com a Saúde. Os *dashboards* construídos, ao reunirem

as principais informações de todos os utilizadores, permitem avaliar os índices de bem-estar cardíaco a um nível global, o que é útil para análises estatísticas. Para os utilizadores-alvo desta página foi construído o conjunto de *dashboards* indicado na **Figura 60**, que apresenta diversos indicadores relativos aos índices globais sob diferentes perspetivas de análise.

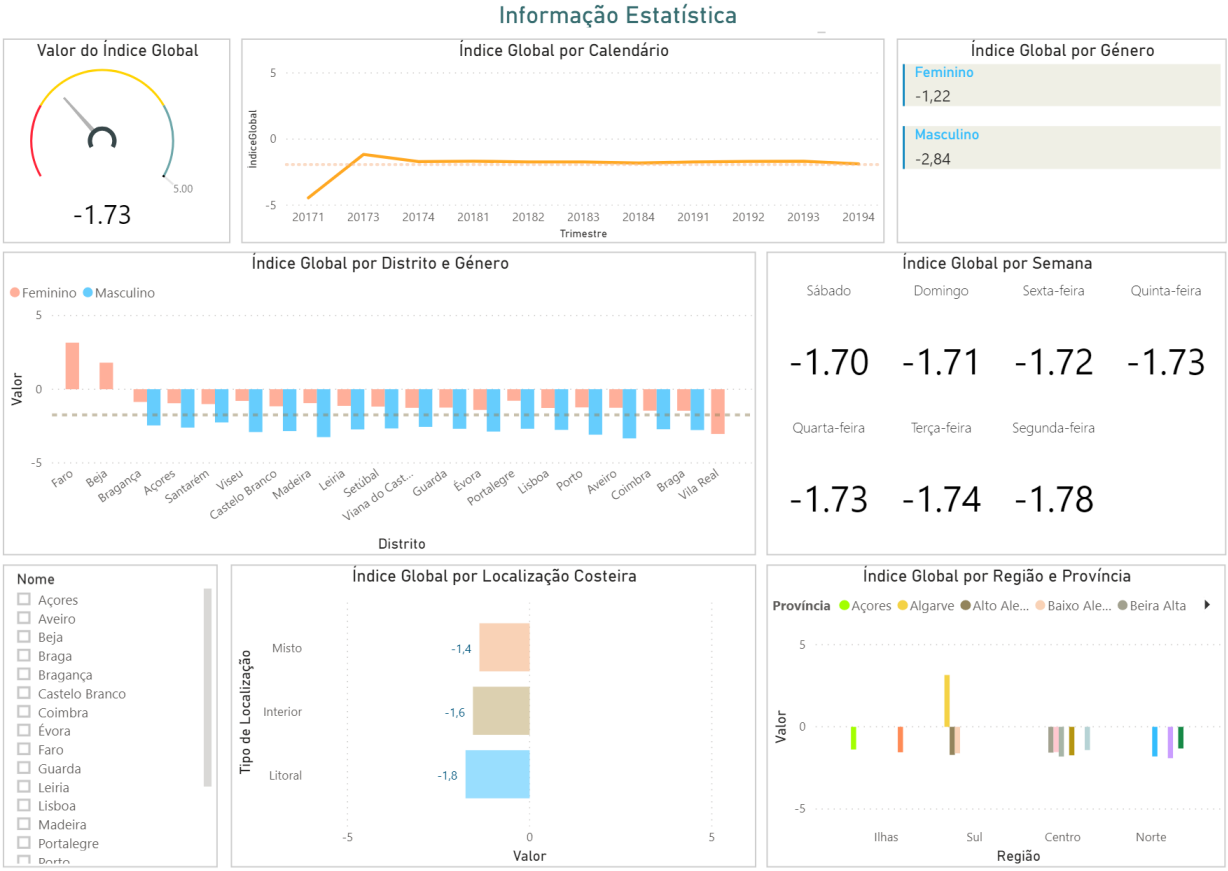


Figura 60 - *Dashboards* para avaliação dos valores de índice globais.

O principal indicador é o que assinala o valor do índice global e, considerando-se todos os registos de todos os utilizadores, em todas as datas, este indicador assume o valor de -1.73. O facto de este valor estar associado a um apontador que remete para uma cor faz com que a sua interpretação se torne mais clara. Neste caso, observa-se que, apesar de ainda não ser um valor crítico (situado na zona vermelha), este valor está próximo desta área e isso reflete que, no caso dos utilizadores em estudo, a sua saúde cardiovascular não está a ser devidamente cuidada e estes devem mudar os seus hábitos, procurando mesmo especialistas para investirem na prevenção.

O índice global também pode ser apresentado de uma forma dependente dos parâmetros em análise, como é o caso do índice global por calendário. Neste caso, de uma forma análoga aos *dashboards* da página anterior, no mesmo gráfico, pode-se navegar pelos índices globais por ano, por trimestre, por mês ou por data. No caso da **Figura 60** exemplificou-se o gráfico do índice global por trimestre, de onde se constata que este índice tem um comportamento

relativamente constante para os trimestres considerados. No que diz respeito aos parâmetros cronológicos, foi também indicado, para cada dia da semana, o valor do índice ordenado de forma decrescente. Em relação a este aspeto, retirou-se que os dias da semana apresentavam, de um modo geral, valores de índices uniformes, ainda que se tenham detetado valores ligeiramente melhores ao sábado e, por oposição, piores resultados à segunda-feira.

Além disso, foi também construído um *dashboard* que permite visualizar de uma forma rápida os índices globais de acordo com o género. De um modo geral, verificou-se que os indivíduos do sexo masculino têm valores significativamente piores do que os do sexo feminino.

Ao nível dos distritos foi colocado um menu para o caso de se pretender analisar os *dashboards* para um local em particular, em vez de se estudarem todos de uma única vez. Ainda neste contexto, criaram-se *dashboards* que possibilitam uma análise do valor global do índice por distrito e por sexo, em simultâneo. Daqui, verificou-se que os utilizadores registados de Faro, Beja e Vila Real são apenas do sexo feminino e, por isso, nota-se que em termos de qualidade, os dados deveriam ser melhorados e passar-se a incluir, de uma forma mais ou menos balanceada, utilizadores de ambos os sexos em cada um dos distritos, para que as conclusões possam ser consideradas válidas. Também ao nível da localização, foi construído um gráfico que possibilita a deteção do valor do índice, consoante os utilizadores residam numa área do país situada no litoral, no interior ou numa zona que seja, simultaneamente, litoral e interior. Desta análise constatou-se que o índice de cada uma destas áreas é semelhante, embora os utilizadores residentes no litoral apresentem piores valores. A última análise efetuada a este nível permitiu relacionar, ao mesmo tempo, a região e a província com o valor do índice, tendo-se verificado que, em todas as províncias, o índice apresentava valores negativos, à exceção do Algarve. Por sua vez, consequentemente, a região sul foi a única do país a conter um registo com um índice positivo.

Caso se pretenda uma análise mais detalhada em relação a aspetos mais particulares, podem-se selecionar e clicar sobre os parâmetros que se pretendem estudar e/ou modificar o nível hierárquico e todos os *dashboards* são imediatamente atualizados de acordo com essa perspetiva de análise, de uma forma interativa.

Como exemplo, ao restringir-se a análise apenas aos dados do distrito de Viana do Castelo, os *dashboards* atualizaram-se de acordo com o que é mostrado na **Figura 61**. Neste caso, optou-se por se visualizarem os índices por mês, em vez de ser observar a sua evolução por trimestre.

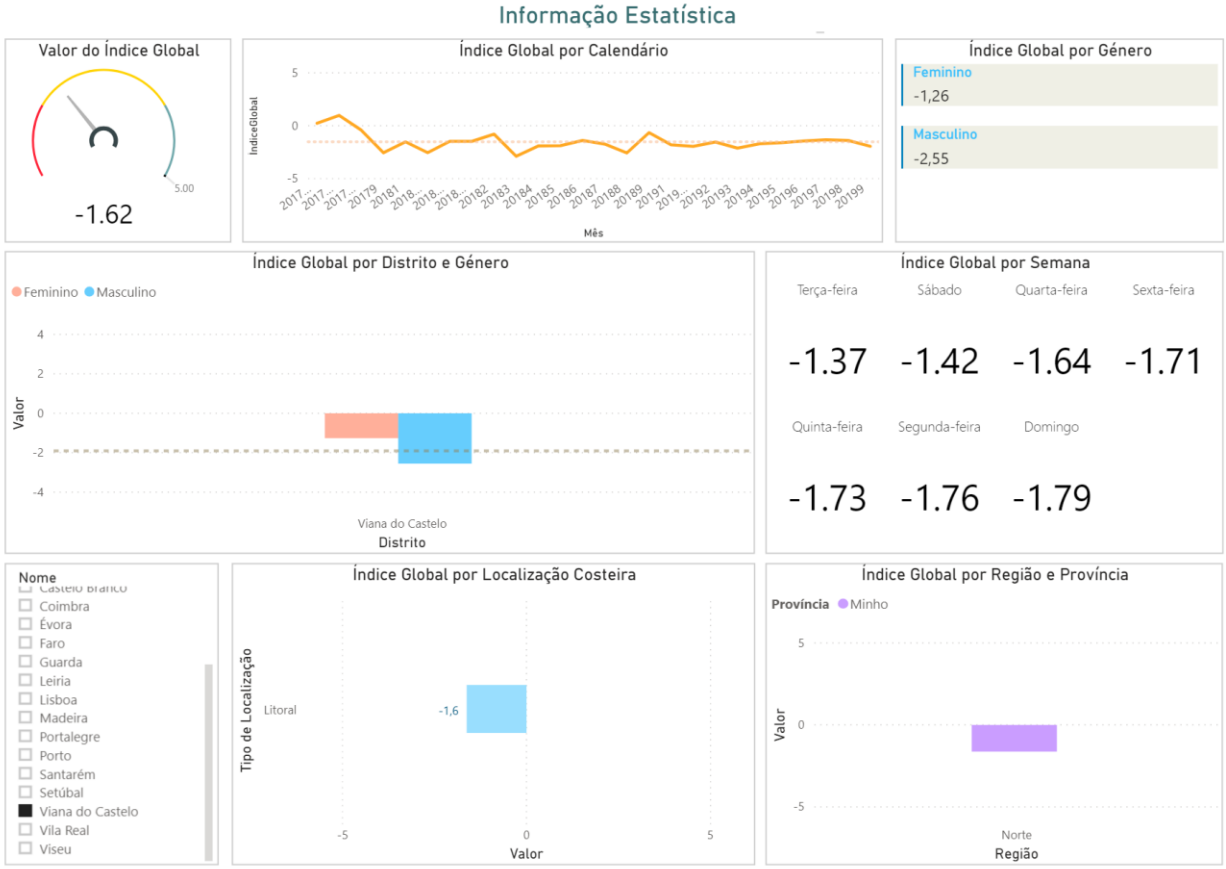


Figura 61 - *Dashboards* para avaliação dos valores de índice globais em Viana do Castelo.

De uma forma genérica, em comparação com os índices globais médios, o valor do índice em Viana do Castelo é ligeiramente superior (-1.62), embora continue relativamente próximo da zona mais crítica, o que mostra que, para os dados em estudo, os habitantes deste distrito também precisam de uma maior sensibilização para o tema das doenças cardiovasculares e devem preocupar-se mais com a sua prevenção, procurando também especialistas. Em termos de género, o índice global é idêntico, em ambos os sexos, ao índice médio de todos os distritos. Por sua vez, em relação aos dias da semana, nota-se que à terça-feira, os habitantes têm comportamentos melhores ao nível da saúde cardiovascular do que nos restantes dias e, pelo lado contrário, o domingo é o dia em que se registam os piores valores.

Além da análise efetuada no Power BI, considerou-se que a disponibilização de um mapa de Portugal, segmentado em distritos, com as correspondentes cores associadas aos valores dos índices em cada um dos distritos, seria uma ferramenta interessante e útil para os decisores. Desta forma, para a sua elaboração, recorreu-se ao *software* QGIS e importou-se a Carta Administrativa Oficial de Portugal, relativa ao ano de 2018, que permite localizar geograficamente todos os concelhos do país. Assim, para se obter o resultado pretendido, numa primeira etapa, foi necessário dissolverem-se os concelhos em distritos. Após isso, importou-

se a tabela com os índices por distrito e, através de uma junção com os distritos da tabela da Carta Administrativa Oficial de Portugal, foram compatibilizados estes índices com a posição geográfica de cada distrito. Posteriormente, definiram-se regras para se associarem os valores dos índices às cores correspondentes (verde, amarelo, laranja e vermelho) e, de forma a colocar-se o valor dos índices visível, especificaram-se rótulos. Na **Figura 62** está apresentado o excerto relativo à zona continental do país, de onde se retira que os índices globais dos distritos são todos negativos, com a exceção dos distritos de Beja e de Faro, que apresentam os melhores valores, sendo que o pior valor se regista em Vila Real.

Deste modo, este mapa serve de complemento aos *dashboards* criados e disponibiliza informação pertinente, de uma forma rápida, concisa e muito elucidativa. Assim, estes mapas estão particularmente orientados para serem integrados como SAD, sobretudo ao nível da Saúde pública. Através da informação disponibilizada por este tipo de mapas podem ser desencadeadas ações que visem promover a Saúde de um modo mais adaptado às necessidades locais como, por exemplo, através da realização de campanhas de sensibilização para as DCV nos distritos que apresentem os piores resultados.

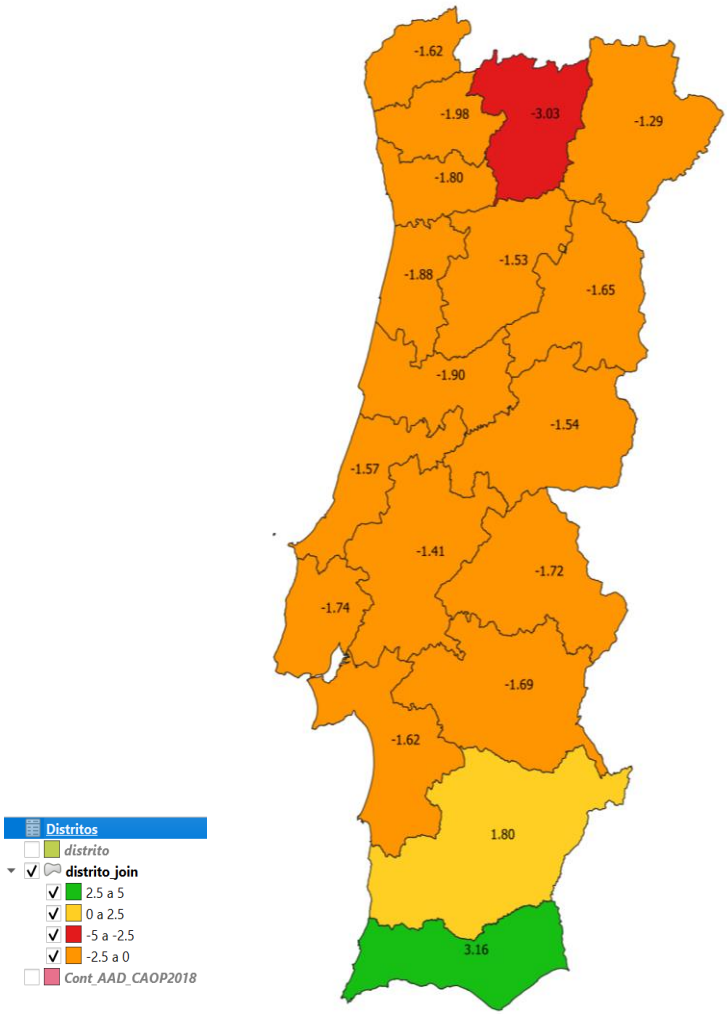


Figura 62 - Mapa de Portugal colorido em função dos valores dos índices de cada distrito.

5.4 *Dashboard* do Histórico dos Utilizadores

Em determinadas situações como, por exemplo, quando se detetam variações abruptas no valor dos índices dos indivíduos, pode ser do interesse dos profissionais a consulta às tabelas de histórico dos utilizadores, para se poder perceber se se registou alguma alteração que pudesse ter influenciado significativamente essa variação. Assim, numa terceira página, foi permitida a consulta, por parte destes profissionais, às tabelas de histórico dos utilizadores com registos modificados, através da construção da tabela ilustrada na **Figura 63**.

Figura 63 - Tabela de histórico do utilizador com *id* 4.

No exemplo da figura, foi selecionada apenas a tabela de histórico relativa ao utilizador com *id* 4 e, através dela, é possível verificarem-se todos os registos passados, com indicação da data de expiração de cada um deles. A título exemplificativo pode ver-se que, no dia 20 de outubro de 2019, as habilitações literárias do utilizador em estudo passaram de mestrado a doutoramento.

6. CONCLUSÕES E TRABALHO FUTURO

Para a concretização do objetivo principal da presente dissertação, ou seja, a elaboração de índices que incentivassem os utilizadores a melhorarem os seus comportamentos, foi necessária uma adequada formulação do índice e foi preciso garantir-se um aspeto final conciso, simples, facilmente entendível e que permitisse a sua monitorização contínua.

O desenvolvimento do sistema proposto, relativo à construção de índices de bem-estar cardíaco, permitiu verificar que os objetivos traçados foram devidamente alcançados.

Em primeiro lugar, concluiu-se que a formulação adotada para o índice era adequada e estava devidamente suportada porque:

- se baseou num sistema de DM que analisou diferentes técnicas e selecionou a melhor, com base em bons valores de acurácia e de sensibilidade. A partir da probabilidade determinada pela técnica escolhida efetuou-se uma correspondência para integração numa escala de valores entre -5 e 5, que garantiu que o número de doentes mal classificados era baixo;
- está em consonância com os simuladores existentes *online*, apresentando valores idênticos ou mais gravosos. No caso das doenças, em geral, a apresentação de valores mais gravosos justifica-se porque é preferível induzir pessoas saudáveis a procurarem profissionais de saúde do que deixar de incentivar pessoas não saudáveis a procurar essa ajuda.

Além disso, concluiu-se também que o índice, apresentado de uma forma gráfica, resume a informação de um modo simples, através de uma escala de cores, facilmente compreendida pela generalidade dos utilizadores. As cores vermelha, laranja e amarela são normalmente identificadas como situações menos positivas, ao invés da verde, que se associa a aspetos positivos. Assim, um utilizador que apresente uma cor "negativa" será induzido a verificar quais os comportamentos que pode adotar com vista a melhorar os seus valores e, com isso, melhorar também a cor do seu índice e a sua saúde cardiovascular. Ainda na vertente dos índices, constatou-se que a escala numérica complementa adequadamente as cores e permite que mesmo aqueles que apresentem um índice na cor verde possam verificar que ainda é possível atingir valores mais elevados.

Através do SDW implementado, verificou-se ainda que a monitorização contínua era possível através da integração dos índices. Com o acompanhamento temporal de cada índice visualizou-se, de uma forma mais abrangente, a sua evolução e, através de uma fórmula de

regressão linear, atenuaram-se os valores mais extremos, criando-se, assim, um índice ponderado que reflete apropriadamente o nível de bem-estar cardíaco. Deste modo, a monitorização temporal apresenta-se como uma mais valia, ao propiciar a determinação de um índice ponderado, que reflete convenientemente, do ponto de vista estatístico, o registo histórico dos índices anteriormente calculados.

Outro dos aspetos positivos que foi possível concluir refere-se à utilização do sistema proposto como um SAD. As decisões podem ser consideradas em três níveis distintos:

- a um nível individual, em que cada utilizador terá ao seu dispor uma ferramenta que lhe permite obter a informação adequada para decidir se pretende mudar o seu estilo de vida, procurando torná-lo mais saudável ou mesmo se se torna prudente a consulta a um profissional de saúde da especialidade;
- ao nível dos profissionais de saúde, em que funcione como um suporte à decisão, complementando diagnósticos, servindo para se efetuarem triagens e se verificarem as necessidades de utilização de exames ou de tratamentos específicos, adequados a cada caso;
- ao nível das decisões de índole geral, realçando geograficamente os índices, de forma a identificarem-se as situações em que seria recomendável um investimento maior em campanhas de sensibilização e servindo como alerta para a urgência de ajustes pontuais aos meios clínicos – humanos e materiais – que sejam necessários incorporar em cada região.

Realça-se ainda que o sistema proposto pode ser efetuado desde o processo de DM até aos *dashboards* finais, de um modo automatizado, em que as atualizações dos indicadores são disponibilizadas em tempo real.

A nível de limitações, uma das dificuldades encontradas prendeu-se com a utilização de dados sintetizados e/ou existentes *online* e que, no futuro, deveriam ser ajustados para dados reais. No entanto, é importante notar que, caso esses dados sejam disponibilizados, o sistema é flexível e consegue-se adaptar, de uma forma relativamente simples, à inserção de novos dados. Além disso, se a existência de dados não fosse uma restrição, no processo de DM, poderiam também incluir-se, em trabalhos futuros, outros atributos para os quais existem, atualmente, evidências de estarem relacionados com as DCV, como é o caso da etnia, do tipo de alimentação, do consumo de álcool e de cafeína e do valor da proteína C reativa.

No que diz respeito a possíveis trabalhos futuros, entendeu-se que o sistema proposto poderia ser complementado pela recolha de dados através de dispositivos eletrónicos móveis,

integrados numa base de dados, com recurso a tecnologia associada à *Internet of Things*. Assim, no futuro, poderia ser testada a integração deste sistema com dados provenientes de utilizadores reais, através de *smartbands* ou de *smartwatches*.

REFERÊNCIAS BIBLIOGRÁFICAS

Abdar, M. *et al.* (2015) 'Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases', *International Journal of Electrical and Computer Engineering*, 5(6), pp. 1569–1576. doi: 10.11591/ijece.v5i6.pp1569-1576.

Andersson, C. *et al.* (2019) '70-Year Legacy of the Framingham Heart Study', *Nature Reviews Cardiology*. Springer US, 1968. doi: 10.1038/s41569-019-0202-5.

Bahrami, B. and Hosseini Shirvani, M. (2015) 'Prediction and Diagnosis of Heart Disease by Data Mining Techniques', *Journal of Multidisciplinary Engineering Science and Technology*, 2(2), pp. 3159–40. Available at: www.jmest.org.

Casters, M., Bouman, R. and Dongen, J. van (2010) *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*. Wiley.

Gerdes, M., Galar, D. and Scholz, D. (2016) 'Automated Parameter Optimization for Feature Extraction for Condition Monitoring', *14th IMEKO TC10 Workshop on Technical Diagnostics 2016: New Perspectives in Measurements, Tools and Techniques for Systems Reliability, Maintainability and Safety*, pp. 452–457.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*. 3rd Editio, *Data Mining: Concepts and Techniques*. 3rd Editio. Morgan Kaufmann Publishers. doi: 10.1016/C2009-0-61819-5.

Hay, S. I. *et al.* (2013) 'Big Data Opportunities for Global Infectious Disease Surveillance', *PLoS Medicine*, 10(4), pp. 1–5. doi: 10.1371/journal.pmed.1001413.

Ho, A. T. S. and Li, S. (2016) *Handbook of Digital Forensics of Multimedia Data and Devices*. John Wiley & Sons.

Jadhav, S. D. and Channe, H. P. (2016) 'Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques', *International Journal of Science and Research (IJSR)*, 5(1), pp. 1842–1845. doi: 10.21275/v5i1.nov153131.

Jankowski, D. and Jackowski, K. (2014) 'Evolutionary Algorithm for Decision Tree Induction', *13th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM)*, pp. 23–32.

John, G. H. and Langley, P. (2013) 'Estimating Continuous Distributions in Bayesian Classifiers', pp. 338–345.

Joseph, S. R., Hlomani, H. and Letsholo, K. (2016) 'Data Mining Algorithms: An Overview', *International Journal of Computers & Technology*, 15(6), pp. 6806–6813. doi: 10.24297/ijct.v15i6.1615.

Kannel, W. B. *et al.* (1961) 'Factors of Risk in the Development of Coronary Heart Disease - Six-Year Follow-up Experience. The Framingham Study.', *Annals of Internal Medicine*, 55(1), pp. 33–50. doi: 10.7326/0003-4819-55-1-33.

Karaolis, M. A. *et al.* (2010) 'Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining with Decision Trees', *IEEE Transactions on Information Technology in Biomedicine*, 14(3), pp. 559–566. doi: 10.1109/TITB.2009.2038906.

Kecman, V. (2005) 'Support Vector Machines – An Introduction', in Wang, L. (ed.) *Support Vector Machines: Theory and Applications*. Springer Berlin Heidelberg, pp. 1–47. doi: 10.1007/10984697_1.

Kim, J. K. and Kang, S. (2017) 'Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis', *Journal of Healthcare Engineering*, 2017. doi: 10.1155/2017/2780501.

Kimball, R. and Caserta, J. (2004) *The data warehouse ETL toolkit : Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley.

Kimball, R. and Ross, M. (2013) *The Data Warehouse Toolkit: the Definitive Guide to Dimensional Modeling*. 3rd Editio. Wiley.

Mahmood, S. S. *et al.* (2014) 'The Framingham Heart Study and the Epidemiology of Cardiovascular Disease: A Historical Perspective', *The Lancet*. Elsevier Ltd, 383(9921), pp. 999–1008. doi: 10.1016/S0140-6736(13)61752-3.

Marrugat, J. *et al.* (2003) 'Estimación del Riesgo Coronario en España Mediante la Ecuación de Framingham Calibrada', *Revista Española de Cardiologia*, 56(3), pp. 253–261. doi: 10.1157/13043951.

Mehta, M., Rissanen, J. and Agrawal, R. (1995) 'MDL-based Decision Tree Pruning', in *Proceedings of Knowledge Discovery in Databases*. Montreal, Canada: AAAI Press, pp. 216–221.

Palaniappan, S. and Awang, R. (2008) 'Intelligent Heart Disease Prediction System Using Data Mining Techniques', *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*, pp. 108–115. doi: 10.1109/AICCSA.2008.4493524.

Patel, N. and Upadhyay, S. (2012) 'Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA', *International Journal of Computer Applications*, 60(12), pp. 20–25. doi: 10.5120/9744-4304.

Pillai, J. J. (2017) *Functional Connectivity, An Issue of Neuroimaging Clinics of North America*. 1st Editio. Elsevier Health Sciences.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. Edited by P. Langley. San Mateo, California: Morgan Kaufmann.

Rajeswari, K., Vaithiyanathan, V. and Neelakantan, T. R. (2012) 'Feature Selection in Ischemic Heart Disease Identification Using Feed Forward Neural Networks', *Procedia Engineering*, 41, pp. 1818–1823. doi: 10.1016/j.proeng.2012.08.109.

Rodrigues, F., Coles, M. and Dye, D. (2012) 'Data Profiling and Scrubbing', in *Pro SQL Server 2012 Integration Services*. Apress, Berkeley, CA, pp. 427–464. doi: 10.1007/978-1-4302-3693-1_12.

S.Dangare, C. and S. Apte, S. (2012) 'Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques', *International Journal of Computer Applications*, 47(10), pp. 44–48. doi: 10.5120/7228-0076.

Santos, V. and Belo, O. (2011) 'No Need to Type Slowly Changing Dimensions', in *Proceedings of the IADIS International Conference Information Systems*, pp. 129–136.

Saravana, N. and Gayathri, V. (2018) 'Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)', *International Journal of Computer Trends and Technology (IJCTT)*, 59(2), pp. 73–80. doi: 10.14445/22312803/ijctt-v59p112.

Sharma, R., Ghosh, A. and Joshi, P. K. (2013) 'Decision Tree Approach for Classification of Remotely Sensed Satellite Data Using Open Source Support', *Journal of Earth System Science*, 122(5), pp. 1237–1247. doi: 10.1007/s12040-013-0339-2.

Sreejith, S., Rahul, S. and Jisha, R. C. (2016) 'A Real Time Patient Monitoring System for Heart Disease Prediction Using Random Forest Algorithm', in *Advances in Signal Processing and Intelligent Recognition Systems*. Springer Verlag, pp. 485–500. doi: 10.1007/978-3-319-28658-7_41.

Stadler, J. G. *et al.* (2016) 'Improving the Efficiency and Ease of Healthcare Analysis Through Use of Data Visualization Dashboards', *Big Data*, 4(2), pp. 129–135. doi: 10.1089/big.2015.0059.

Taneja, A. (2013) 'Heart Disease Prediction System Using Data Mining Techniques', *Oriental Journal Of Computer Science & Technology*, 6, pp. 457–466.

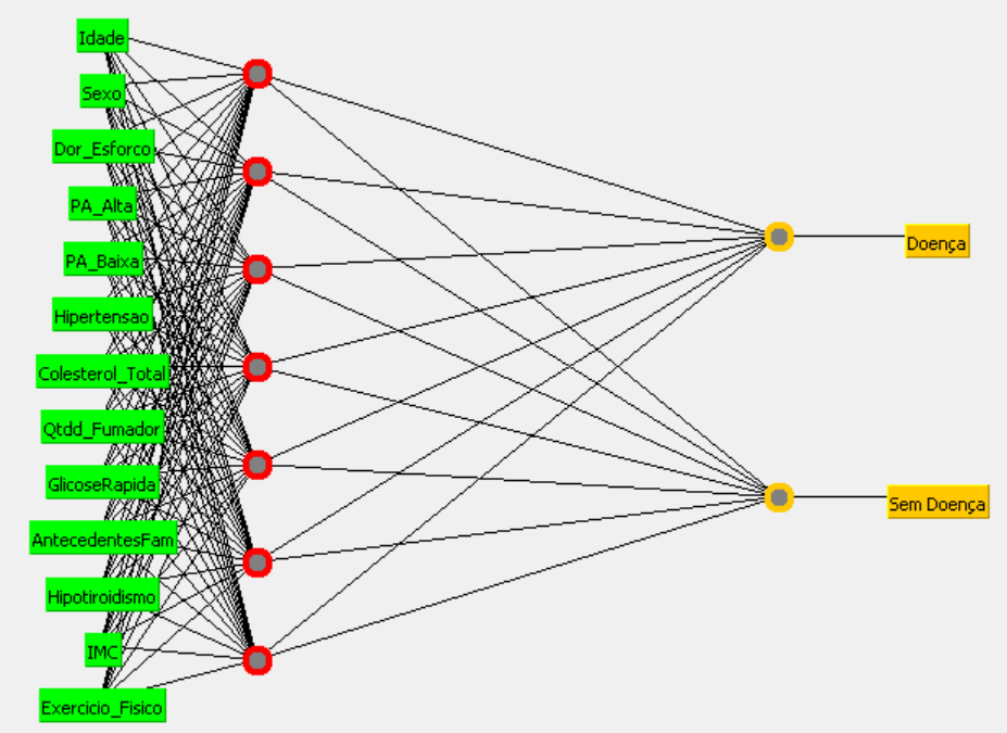
Wilson, P. W. F. *et al.* (1998) 'Prediction of Coronary Heart Disease Using Risk Factor Categories', *Circulation*, 97(18), pp. 1837–1847. doi: 10.1161/01.CIR.97.18.1837.

Witten, I. H. *et al.* (2016) *The WEKA Workbench*. 4th Editio. Morgan Kaufmann. doi: 10.1016/b978-0-12-804291-5.00024-6.

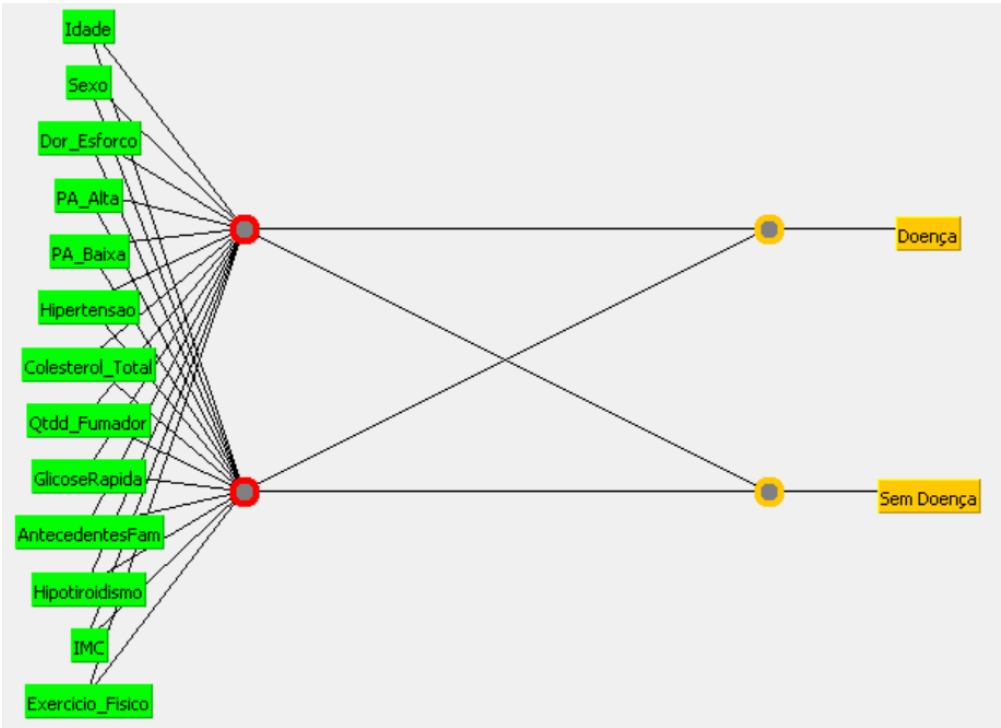
Witten, I. H. *et al.* (2017) *Data Mining: Practical Machine Learning Tools and Techniques*. 4th Edi. Morgan Kaufmann Publishers. doi: 0120884070, 9780120884070.

ANEXO I – ARQUITETURA DAS REDES NEURONAIS MLP (EXEMPLO CENÁRIO I)

Nº de Camadas Escondidas = 1 com "*a*" Neurónios:



Nº de Camadas Escondidas = 1 com "*o*" Neurónios:



ANEXO II – SIMULAÇÕES EFETUADAS PARA A OTIMIZAÇÃO DOS PARÂMETROS DAS TÉCNICAS DE DM (CENÁRIO I)

J48:

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.10	**2**	**F**	**F**	**3**	**V**	75.0014	14.7426	36
					F	74.5494	14.1644	1:03
				6	**V**	75.0014	14.7426	33
					F	74.5494	14.1644	1:03
			V	**3**	**V**	71.2682	15.5017	31
					F	66.8098	16.3757	58
				6	**V**	71.2682	15.5017	34
					F	66.8098	16.3757	1:01
		V	**F**	**3**	**V**	74.6059	14.2755	23
					F	74.1331	14.2322	39
				6	**V**	74.6096	14.3075	31
					F	73.3721	14.4055	53
	4	**F**	**F**	**3**	**V**	75.0315	14.8048	21
					F	74.5851	14.2378	44
				6	**V**	75.0315	14.8048	26
					F	74.5851	14.2378	48
			V	**3**	**V**	72.1855	15.1476	26
					F	68.1019	15.9067	40
				6	**V**	72.1855	15.1476	24
					F	68.1019	15.9067	43
		V	**F**	**3**	**V**	74.781	14.2378	18
					F	74.4156	14.0947	29
				6	**V**	74.668	14.3716	18
					F	73.7018	14.5091	38
	6	**F**	**F**	**3**	**V**	75.0504	14.9649	21
					F	74.7547	14.2698	35
				6	**V**	75.0504	14.9649	18
					F	74.7547	14.2698	34
			V	**3**	**V**	72.824	14.8255	20
					F	69.3451	15.4640	33
				6	**V**	72.824	14.8255	19
					F	69.3451	15.4640	34
		V	**F**	**3**	**V**	74.7886	14.5128	13
					F	74.5155	14.3037	22
				6	**V**	74.8244	14.1700	14
					F	74.0653	14.3960	29
	10	**F**	**F**	**3**	**V**	75.0862	15.0685	12
					F	74.8375	14.4450	20
				6	**V**	75.0862	15.0685	14
					F	74.8375	14.4450	20
			V	**3**	**V**	73.6961	14.4017	12
					F	71.0459	15.0798	18

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.10	**10**	**F**	**V**	**6**	**V**	73.6961	14.4017	11
					F	71.0459	15.0798	20
		V	**F**	**3**	**V**	75.0447	14.4525	8
					F	74.7923	14.2830	14
				6	**V**	75.0899	14.2755	9
					F	74.4458	14.2849	17
	20	**F**	**F**	**3**	**V**	75.1088	15.0195	8
					F	74.9619	14.5844	13
				6	**V**	75.1088	15.0195	8
					F	74.9619	14.5844	14
			V	**3**	**V**	74.4929	14.1323	8
					F	73.2139	14.2981	13
				6	**V**	74.4929	14.1323	8
					F	73.2139	14.2981	13
		V	**F**	**3**	**V**	75.1766	14.6051	5
					F	75.0749	14.3885	8
				6	**V**	75.1408	14.4017	6
					F	74.6699	14.3019	10
0.25	**2**	**F**	**F**	**3**	**V**	74.4401	14.2924	37
					F	72.08	14.7765	1:02
				6	**V**	74.4401	14.2924	33
					F	72.08	14.7765	1:04
			V	**3**	**V**	71.2682	15.5017	36
					F	66.8098	16.3757	1:04
				6	**V**	71.2682	15.5017	37
					F	66.8098	16.3757	50
		V	**F**	**3**	**V**	74.6059	14.2755	25
					F	74.1331	14.2322	42
				6	**V**	74.6096	14.3075	31
					F	73.3721	14.4055	51
	4	**F**	**F**	**3**	**V**	74.7019	14.2906	26
					F	73.1405	14.4789	49
				6	**V**	74.7019	14.2906	24
					F	73.1405	14.4789	45
			V	**3**	**V**	72.1855	15.1476	26
					F	68.1019	15.9067	44
				6	**V**	72.1855	15.1476	25
					F	68.1019	15.9067	40
		V	**F**	**3**	**V**	74.781	14.2378	16
					F	74.4156	14.0947	26
				6	**V**	74.668	14.3716	21
					F	73.7018	14.5091	38
	6	**F**	**F**	**3**	**V**	74.8375	14.2378	21
					F	73.9391	14.0476	37
				6	**V**	74.8375	14.2378	18
					F	73.9391	14.0476	34
			V	**3**	**V**	72.824	14.8255	20
					F	69.3451	15.4640	31
				6	**V**	72.824	14.8255	20

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.25	**6**	**F**	**V**	**6**	**F**	69.3451	15.4640	35
		V	**F**	**3**	**V**	74.7886	14.5128	12
					F	74.5155	14.3037	21
				6	**V**	74.8244	14.1700	16
					F	74.0653	14.3960	27
	10	**F**	**F**	**3**	**V**	75.0089	14.2698	12
					F	74.2254	14.0890	18
				6	**V**	75.0089	14.2698	12
					F	74.2254	14.0890	19
			V	**3**	**V**	73.6961	14.4017	12
					F	71.0459	15.0798	19
				6	**V**	73.6961	14.4017	12
					F	71.0459	15.0798	20
		V	**F**	**3**	**V**	75.0447	14.4525	7
					F	74.7923	14.2830	13
				6	**V**	75.0899	14.2755	10
					F	74.4458	14.2849	16
	20	**F**	**F**	**3**	**V**	75.1333	14.4450	8
					F	74.6379	14.1135	13
				6	**V**	75.1333	14.4450	8
					F	74.6379	14.1135	12
			V	**3**	**V**	74.4929	14.1323	8
					F	73.2139	14.2981	12
				6	**V**	74.4929	14.1323	7
					F	73.2139	14.2981	12
		V	**F**	**3**	**V**	75.1766	14.6051	5
					F	75.0749	14.3885	9
				6	**V**	75.1408	14.4017	6
					F	74.6699	14.3019	12
0.40	**2**	**F**	**F**	**3**	**V**	72.9559	14.7765	34
					F	68.9081	15.7334	1:03
				6	**V**	72.9559	14.7765	25
					F	68.9081	15.7334	52
			V	**3**	**V**	71.2682	15.5017	27
					F	66.8098	16.3757	48
				6	**V**	71.2682	15.5017	24
					F	66.8098	16.3757	46
		V	**F**	**3**	**V**	74.6059	14.2755	21
					F	74.1331	14.2322	31
				6	**V**	74.6096	14.3075	23
					F	73.3721	14.4055	40
	4	**F**	**F**	**3**	**V**	73.7074	14.6183	21
					F	70.5807	15.1721	33
				6	**V**	73.7074	14.6183	20
					F	70.5807	15.1721	38
			V	**3**	**V**	72.1855	15.1476	19
					F	68.1019	15.9067	33
				6	**V**	72.1855	15.1476	19
					F	68.1019	15.9067	33

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.40	**4**	**V**	**F**	**3**	**V**	74.781	14.2378	13
					F	74.4156	14.0947	22
				6	**V**	74.668	14.3716	16
					F	73.7018	14.5091	28
	6	**F**	**F**	**3**	**V**	74.086	14.3471	16
					F	71.837	14.8519	31
				6	**V**	74.086	14.3471	17
					F	71.837	14.8519	29
			V	**3**	**V**	72.824	14.8255	16
					F	69.3451	15.4640	27
				6	**V**	72.824	14.8255	14
					F	69.3451	15.4640	27
		V	**F**	**3**	**V**	74.7886	14.5128	11
					F	74.5155	14.3037	18
				6	**V**	74.8244	14.1700	12
					F	74.0653	14.3960	23
	10	**F**	**F**	**3**	**V**	74.6115	14.0683	13
					F	72.7317	14.5787	21
				6	**V**	74.6115	14.0683	12
					F	72.7317	14.5787	20
			V	**3**	**V**	73.6961	14.4017	12
					F	71.0459	15.0798	18
				6	**V**	73.6961	14.4017	12
					F	71.0459	15.0798	20
		V	**F**	**3**	**V**	75.0447	14.4525	7
					F	74.7923	14.2830	12
				6	**V**	75.0899	14.2755	9
					F	74.4458	14.2849	17
	20	**F**	**F**	**3**	**V**	74.911	14.1794	8
					F	74.1632	14.1116	14
				6	**V**	74.911	14.1794	8
					F	74.1632	14.1116	14
			V	**3**	**V**	74.4929	14.1323	7
					F	73.2139	14.2981	13
				6	**V**	74.4929	14.1323	8
					F	73.2139	14.2981	13
		V	**F**	**3**	**V**	75.1766	14.6051	5
					F	75.0749	14.3885	9
				6	**V**	75.1408	14.4017	7
					F	74.6699	14.3019	11

Random Forest:

Nº Árvores	*Profund. Máx*	*Nº Atributos*	*BreakTies*	*Accuracy*	*FN (%)*	*Tempo*
20	**0**	**0**	**V**	73.7545	13.2885	15
			F	73.7601	13.0568	16
		5	**V**	73.4964	13.2791	18
			F	73.3778	13.4034	17
		10	**V**	73.1084	13.4204	28
			F	73.0783	13.3827	28
	10	**0**	**V**	75.4723	14.9479	9
			F	75.4516	14.9498	9
		5	**V**	75.3555	15.0082	11
			F	75.4384	14.9611	11
		10	**V**	75.2971	14.9988	19
			F	75.2764	14.9969	18
	25	**0**	**V**	74.0182	13.8291	17
			F	74.0898	13.6577	15
		5	**V**	74.0634	13.6822	20
			F	73.9542	13.8272	18
		10	**V**	73.5398	13.8686	29
			F	73.4776	13.9063	30
50	**0**	**0**	**V**	74.4363	13.7424	42
			F	74.3083	13.5748	39
		5	**V**	74.2649	13.6370	47
			F	74.1858	13.7029	43
		10	**V**	73.8016	13.8573	1:18
			F	73.8336	13.8555	1:15
	10	**0**	**V**	75.4742	14.9573	22
			F	75.4798	14.9423	22
		5	**V**	75.5929	14.8613	26
			F	75.591	14.8462	27
		10	**V**	75.4346	14.8745	47
			F	75.4064	14.8537	49
	25	**0**	**V**	74.6059	13.9666	36
			F	74.523	13.9591	36
		5	**V**	74.5286	13.9346	46
			F	74.4834	13.9948	53
		10	**V**	74.2348	14.0608	1:22
			F	74.0125	14.1154	1:10

Nº Árvores	*Profund. Máx*	*Nº Atributos*	*BreakTies*	*Accuracy (%)*	*FN (%)*	*Tempo*
80	**0**	**0**	**V**	74.5625	13.7858	1:03
			F	74.4759	13.7462	1:05
		5	**V**	74.4589	13.7462	1:18
			F	74.3817	13.7895	1:16
		10	**V**	73.9259	13.9534	1:47
			F	73.8995	13.9609	1:52
	10	**0**	**V**	75.5175	14.9837	37
			F	75.6098	14.8989	36
		5	**V**	75.6004	14.8481	39
			F	75.5684	14.8556	40
		10	**V**	75.4064	14.9291	1:12
			F	75.4516	14.8952	1:15
	25	**0**	**V**	74.7886	13.9854	57
			F	74.7358	13.9289	58
		5	**V**	74.5682	14.0325	1:10
			F	74.6661	13.9572	1:07
		10	**V**	74.2423	14.0758	1:58
			F	74.0031	14.2246	1:53
100	**0**	**0**	**V**	74.5588	13.8611	1:19
			F	74.5173	13.8329	1:32
		5	**V**	74.4834	13.7839	1:31
			F	74.4363	13.8103	1:29
		10	**V**	74.0615	13.9741	2:34
			F	73.8788	14.0269	2:29
	10	**0**	**V**	75.5476	14.9536	41
			F	75.5815	14.9291	42
		5	**V**	75.6079	14.8424	50
			F	75.5476	14.8650	53
		10	**V**	75.4327	14.9178	1:27
			F	75.4648	14.9084	1:29
	25	**0**	**V**	74.7264	14.0156	1:14
			F	74.717	13.9930	1:17
		5	**V**	74.6134	14.0043	1:35
			F	74.5512	14.0513	1:39
		10	**V**	74.2894	14.1210	2:26
			F	74.0935	14.1983	2:28

Naive Bayes:

Kernel	*Discretização Supervisionada*	*Accuracy (%)*	*FN (%)*	*Tempo*
V	**F**	73.7846	17.0707	0
F	**V**	73.8788	16.6262	1
	F	72.6093	17.9277	0

KNN:

k	*CrossValidate*	*Distância Ponder.*	*Algoritmo Pesquisa VMP*	*Accuracy (%)*	*FN (%)*	*Tempo*
10	**V**	**Não**	***LinearNNSearch***	71.9953	14.9197	15:58
			BallTree	71.9953	14.9197	7:21
			FilteredNSearch	65.7513	17.6885	10:39
			KDTree	71.9953	14.9197	2:09
		1/dist	***LinearNNSearch***	71.9387	15.7560	17:02
			BallTree	71.9387	15.7560	6:50
			FilteredNSearch	65.7513	17.6885	11:37
			KDTree	71.9387	15.7560	2:11
		1-dist	***LinearNNSearch***	72.0292	15.7352	18:44
			BallTree	72.0292	15.7352	6:27
			FilteredNSearch	65.7513	17.6885	10:57
			KDTree	72.0292	15.7352	1:57
	F	**Não**	***LinearNNSearch***	71.9858	13.5013	1:27
			BallTree	71.9858	13.5013	37
			FilteredNSearch	71.9858	13.5013	1:27
			KDTree	71.9858	13.5013	14
		1/dist	***LinearNNSearch***	72.0235	15.7371	1:54
			BallTree	72.0235	15.7371	38
			FilteredNSearch	72.0235	15.7371	1:11
			KDTree	72.0235	15.7371	13
		1-dist	***LinearNNSearch***	72.1139	15.6863	2:12
			BallTree	72.1139	15.6863	36
			FilteredNSearch	72.1139	15.6863	1:15
			KDTree	72.1139	15.6863	13
20	**V**	**Não**	***LinearNNSearch***	72.6978	15.4150	20:42
			BallTree	72.6978	15.4150	7:35
			FilteredNSearch	65.7513	17.6885	11:10
			KDTree	72.6997	15.4132	2:36
		1/dist	***LinearNNSearch***	72.7807	15.9293	17:19
			BallTree	72.7807	15.9293	7:41
			FilteredNSearch	65.7513	17.6885	11:19
			KDTree	72.7788	15.9293	2:53
		1-dist	***LinearNNSearch***	72.7487	15.9989	3:47
			BallTree	72.7487	15.9989	6:40
			FilteredNSearch	65.7513	17.6885	11:34
			KDTree	72.7506	15.9971	2:58
	F	**Não**	***LinearNNSearch***	72.7694	14.8123	2:01
			BallTree	72.7694	14.8123	44
			FilteredNSearch	72.7694	14.8123	1:08
			KDTree	72.7713	14.8104	19
		1/dist	***LinearNNSearch***	72.7487	15.9443	1:55
			BallTree	72.7487	15.9443	48
			FilteredNSearch	72.7487	15.9443	1:07
			KDTree	72.7506	15.9424	16
		1-dist	***LinearNNSearch***	72.7487	15.9989	2:07
			BallTree	72.7487	15.9989	43
			FilteredNSearch	72.7487	15.9989	1:06
			KDTree	72.7506	15.9971	17

k	*CrossValidate*	*Distância Ponder.*	*Algoritmo Pesquisa VMP*	*Accuracy (%)*	*FN (%)*	*Tempo*
50	**V**	**Não**	***LinearNNSearch***	72.9276	15.7070	19:06
			BallTree	72.9276	15.7070	7:56
			FilteredNSearch	65.7513	17.6885	12:51
			KDTree	72.9295	15.7070	3:52
		1/dist	***LinearNNSearch***	72.9596	16.0705	17:48
			BallTree	72.9596	16.0705	9:09
			FilteredNSearch	65.7513	17.6885	11:43
			KDTree	72.9596	16.0686	3:57
		1-dist	***LinearNNSearch***	72.873	16.1120	18:36
			BallTree	72.873	16.1120	9:15
			FilteredNSearch	65.7513	17.6885	11:46
			KDTree	72.873	16.1120	3:48
	F	**Não**	***LinearNNSearch***	72.792	16.0197	1:37
			BallTree	72.792	16.0197	53
			FilteredNSearch	72.792	16.0197	1:13
			KDTree	72.7958	16.0197	24
		1/dist	***LinearNNSearch***	72.9333	16.3286	1:55
			BallTree	72.9333	16.3286	54
			FilteredNSearch	72.9333	16.3286	1:10
			KDTree	72.9351	16.3286	24
		1-dist	***LinearNNSearch***	72.7901	16.4435	1:45
			BallTree	72.7901	16.4435	53
			FilteredNSearch	72.7901	16.4435	1:13
			KDTree	72.7939	16.4435	25

MultiLayer Perceptron:

Camadas Escond.	*Tempo Treino*	*Nominal para Bin*	*Decay*	*LR*	*Momentum*	*Acuraccy (%)*	*FN (%)*	*Tempo*
a	**100**	**V**	**V**	**0.1**	**0.05**	74.7792	14.6842	58
					0.1	74.7848	14.6786	57
					0.2	74.847	14.6635	58
				0.2	**0.05**	74.9788	14.6918	58
					0.1	75.041	14.6579	59
					0.2	75.1012	14.6051	58
				0.3	**0.05**	75.0768	14.5731	57
					0.1	75.1351	14.5467	58
					0.2	75.1464	14.5204	58
			F	**0.1**	**0.05**	75.25	14.1079	1:02
					0.1	75.2387	14.0852	1:01
					0.2	75.1634	14.0758	58
				0.2	**0.05**	75.1163	14.0212	57
					0.1	74.9807	14.0947	59
					0.2	74.975	13.9346	1:00
				0.3	**0.05**	74.8244	14.0513	1:02
					0.1	74.7208	14.0739	58
					0.2	74.5946	14.0193	59
		F	**V**	**0.1**	**0.05**	74.6944	14.8707	42
					0.1	74.6831	14.8971	43
					0.2	74.7	14.9178	46
				0.2	**0.05**	74.9261	14.8650	42
					0.1	74.9185	14.8575	44
					0.2	74.8884	14.8311	44
				0.3	**0.05**	74.9694	14.8180	42
					0.1	74.9656	14.7916	45
					0.2	74.9882	14.7803	43
			F	**0.1**	**0.05**	75.1238	13.9383	44
					0.1	75.0749	13.9214	43
					0.2	74.8658	14.0438	44
				0.2	**0.05**	74.8206	14.0269	44
					0.1	74.7415	13.9289	44
					0.2	74.7	14.1323	42
				0.3	**0.05**	74.5569	14.0909	44
					0.1	74.4608	14.1719	44
					0.2	74.3327	14.0193	44
	300	**V**	**V**	**0.1**	**0.05**	74.8357	14.6371	3:05
					0.1	74.8696	14.6428	3:04
					0.2	74.9185	14.6428	2:57
				0.2	**0.05**	75.0843	14.5693	3:05
					0.1	75.0975	14.5957	3:01
					0.2	75.0937	14.5618	3:04
				0.3	**0.05**	75.1521	14.5279	3:04
					0.1	75.2161	14.5034	3:03
					0.2	75.2124	14.4507	3:07
			F	**0.1**	**0.05**	75.3197	14.0834	3:01
					0.1	75.2048	14.0947	2:56

Camadas Escond.	*Tempo Treino*	*Nominal para Bin*	*Decay*	*LR*	*Momentum*	*Acuraccy (%)*	*FN (%)*	*Tempo*
a	**300**	**V**	**F**	**0.1**	**0.2**	75.1879	14.0966	2:55
				0.2	**0.05**	75.0504	13.9948	3:00
					0.1	74.9732	14.1305	3:06
					0.2	74.9656	13.9666	3:00
				0.3	**0.05**	74.6548	13.8837	2:58
					0.1	74.7415	13.7255	3:04
					0.2	74.6172	13.8216	2:58
		F	**V**	**0.1**	**0.05**	74.7396	14.8669	2:09
					0.1	74.7848	14.8650	2:10
					0.2	74.8507	14.8349	2:11
				0.2	**0.05**	75.0127	14.7426	2:10
					0.1	75.0146	14.7313	2:11
					0.2	75.0127	14.6974	2:14
				0.3	**0.05**	75.0127	14.7388	2:09
					0.1	75.1069	14.6428	2:11
					0.2	75.0786	14.6918	2:11
			F	**0.1**	**0.05**	75.0843	14.0137	2:10
					0.1	75.0692	13.9967	2:11
					0.2	74.8884	14.0438	2:10
				0.2	**0.05**	74.8055	14.0852	2:10
					0.1	74.6982	13.9930	2:10
					0.2	74.6304	14.2246	2:11
				0.3	**0.05**	74.5701	14.0438	2:10
					0.1	74.3874	14.2529	2:10
					0.2	74.3648	14.0043	2:11
o	**100**	**V**	**V**	**0.1**	**0.05**	74.3893	14.6710	19
					0.1	74.4269	14.6447	19
					0.2	74.4495	14.5900	19
				0.2	**0.05**	74.3949	14.5147	19
					0.1	74.3893	14.5185	19
					0.2	74.3968	14.5015	19
				0.3	**0.05**	74.3855	14.4262	19
					0.1	74.3817	14.4149	20
					0.2	74.312	14.5109	19
			F	**0.1**	**0.05**	73.8054	14.5467	19
					0.1	73.7809	14.5693	19
					0.2	73.666	14.6409	19
				0.2	**0.05**	73.3514	14.4959	19
					0.1	73.3458	14.4996	19
					0.2	73.2064	14.7370	19
				0.3	**0.05**	73.0387	14.5712	19
					0.1	73.0124	14.5335	19
					0.2	72.9088	14.8349	19
		F	**V**	**0.1**	**0.05**	74.3987	14.8161	18
					0.1	74.41	14.7445	18
					0.2	74.4137	14.7633	18
				0.2	**0.05**	74.4137	14.5938	19
					0.1	74.4382	14.5787	18
					0.2	74.4589	14.5599	18

Camadas Escond.	*Tempo Treino*	*Nominal para Bin*	*Decay*	*LR*	*Momentum*	*Acuraccy (%)*	*FN (%)*	*Tempo*
o	**100**	**F**	**V**	**0.3**	**0.05**	74.4288	14.5957	18
					0.1	74.4288	14.5882	18
					0.2	74.4269	14.5712	18
			F	**0.1**	**0.05**	73.8656	14.4130	18
					0.1	73.8562	14.3753	18
					0.2	73.7545	14.4714	18
				0.2	**0.05**	73.4023	14.5863	18
					0.1	73.3514	14.5919	18
					0.2	73.2516	14.5844	18
				0.3	**0.05**	73.0369	14.3904	17
					0.1	73.0105	14.3527	17
					0.2	72.9276	14.3452	18
	300	**V**	**V**	**0.1**	**0.05**	74.442	14.5976	57
					0.1	74.4363	14.6108	56
					0.2	74.4778	14.5806	59
				0.2	**0.05**	74.4382	14.5298	58
					0.1	74.4627	14.5241	57
					0.2	74.4816	14.5241	56
				0.3	**0.05**	74.5475	14.5015	55
					0.1	74.555	14.4996	55
					0.2	74.5117	14.5580	56
			F	**0.1**	**0.05**	73.8016	14.4883	56
					0.1	73.8167	14.4902	57
					0.2	73.6961	14.6126	56
				0.2	**0.05**	73.3608	14.4808	56
					0.1	73.3571	14.4751	56
					0.2	73.2083	14.7181	56
				0.3	**0.05**	73.0482	14.5693	57
					0.1	73.018	14.5354	55
					0.2	72.9125	14.8198	56
		F	**V**	**0.1**	**0.05**	74.4476	14.7313	52
					0.1	74.4665	14.6710	52
					0.2	74.4684	14.6692	53
				0.2	**0.05**	74.3968	14.6390	53
					0.1	74.4006	14.6484	54
					0.2	74.4024	14.6616	55
				0.3	**0.05**	74.425	14.6466	52
					0.1	74.442	14.6371	58
					0.2	74.4476	14.6371	54
			F	**0.1**	**0.05**	73.9014	14.3527	56
					0.1	73.8637	14.3358	58
					0.2	73.7564	14.4394	58
				0.2	**0.05**	73.3778	14.5919	56
					0.1	73.3382	14.5938	57
					0.2	73.2554	14.5787	54
				0.3	**0.05**	73.0406	14.3847	54
					0.1	73.003	14.3527	54
					0.2	72.9182	14.3452	54

ANEXO III – SIMULAÇÕES EFETUADAS PARA A OTIMIZAÇÃO DOS PARÂMETROS DAS TÉCNICAS DE DM (CENÁRIO II)

J48:

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.10	**2**	**F**	**F**	**3**	**V**	74.96	15.8388	13
					F	74.4571	14.8989	47
				6	**V**	74.96	15.8388	15
					F	74.4571	14.8989	45
			V	**3**	**V**	72.9351	15.4923	15
					F	66.8908	16.5263	57
				6	**V**	72.9351	15.4923	14
					F	66.8908	16.5263	44
		V	**F**	**3**	**V**	74.7245	14.7483	10
					F	73.6867	14.3508	32
				6	**V**	74.5588	14.9686	12
					F	73.1932	14.3113	38
	4	**F**	**F**	**3**	**V**	74.9185	15.8483	10
					F	74.5362	14.9460	27
				6	**V**	74.9185	15.8483	10
					F	74.5362	14.9460	30
			V	**3**	**V**	73.3439	15.3209	9
					F	68.134	16.0799	29
				6	**V**	73.3439	15.3209	9
					F	68.134	16.0799	35
		V	**F**	**3**	**V**	74.8357	14.7238	7
					F	73.8355	14.5317	25
				6	**V**	74.7396	14.7822	8
					F	73.6245	14.3320	29
	6	**F**	**F**	**3**	**V**	74.9054	15.8388	8
					F	74.6341	14.9668	24
				6	**V**	74.9054	15.8388	8
					F	74.6341	14.9668	26
			V	**3**	**V**	73.7281	15.1589	7
					F	69.7161	15.4923	27
				6	**V**	73.7281	15.1589	8
					F	69.7161	15.4923	23
		V	**F**	**3**	**V**	74.8507	14.8161	3
					F	74.1802	14.6579	18
				6	**V**	74.7038	14.7426	7
					F	73.8788	14.3132	21
	10	**F**	**F**	**3**	**V**	74.8451	15.7503	6
					F	74.7716	14.8048	19
				6	**V**	74.8451	15.7503	7
					F	74.7716	14.8048	19
			V	**3**	**V**	74.0107	15.0195	6
					F	71.6732	14.8368	18

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.10	**10**	**F**	**V**	**6**	**V**	74.0107	15.0195	7
					F	71.6732	14.8368	20
		V	**F**	**3**	**V**	74.9336	14.7803	4
					F	74.5155	14.7275	12
				6	**V**	74.8187	14.6955	5
					F	74.1576	14.2811	14
	20	**F**	**F**	**3**	**V**	74.9148	15.7183	5
					F	74.8319	15.1758	10
				6	**V**	74.9148	15.7183	5
					F	74.8319	15.1758	10
			V	**3**	**V**	74.4646	14.8462	4
					F	73.325	14.3150	10
				6	**V**	74.4646	14.8462	4
					F	73.325	14.3150	10
		V	**F**	**3**	**V**	75.0429	14.6692	3
					F	74.7321	14.6692	7
				6	**V**	74.975	14.6484	3
					F	74.685	14.4940	9
0.25	**2**	**F**	**F**	**3**	**V**	74.6228	15.2794	13
					F	72.8862	14.7709	44
				6	**V**	74.6228	15.2794	14
					F	72.8862	14.7709	52
			V	**3**	**V**	72.9351	15.4923	13
					F	66.8908	16.5263	45
				6	**V**	72.9351	15.4923	15
					F	66.8908	16.5263	51
		V	**F**	**3**	**V**	74.7245	14.7483	12
					F	73.6867	14.3508	32
				6	**V**	74.5588	14.9686	12
					F	73.1932	14.3113	41
	4	**F**	**F**	**3**	**V**	74.781	15.1306	11
					F	73.3495	14.5599	36
				6	**V**	74.781	15.1306	10
					F	73.3495	14.5599	36
			V	**3**	**V**	73.3439	15.3209	9
					F	68.134	16.0799	29
				6	**V**	73.3439	15.3209	10
					F	68.134	16.0799	30
		V	**F**	**3**	**V**	74.8357	14.7238	8
					F	73.8355	14.5317	23
				6	**V**	74.7396	14.7822	8
					F	73.6245	14.3320	28
	6	**F**	**F**	**3**	**V**	74.8394	15.1344	7
					F	73.828	14.5185	24
				6	**V**	74.8394	15.1344	8
					F	73.828	14.5185	24
			V	**3**	**V**	73.7281	15.1589	7
					F	69.7161	15.4923	25
				6	**V**	73.7281	15.1589	8

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.25	**6**	**F**	**V**	**6**	**F**	69.7161	15.4923	24
		V	**F**	**3**	**V**	74.8507	14.8161	7
					F	74.1802	14.6579	17
				6	**V**	74.7038	14.7426	7
					F	73.8788	14.3132	22
	10	**F**	**F**	**3**	**V**	74.862	15.1438	5
					F	74.2593	14.4431	18
				6	**V**	74.862	15.1438	6
					F	74.2593	14.4431	19
			V	**3**	**V**	74.0107	15.0195	6
					F	71.6732	14.8368	16
				6	**V**	74.0107	15.0195	6
					F	71.6732	14.8368	19
		V	**F**	**3**	**V**	74.9336	14.7803	4
					F	74.5155	14.7275	10
				6	**V**	74.8187	14.6955	5
					F	74.1576	14.2811	14
	20	**F**	**F**	**3**	**V**	74.8959	15.2832	4
					F	74.6172	14.3866	11
				6	**V**	74.8959	15.2832	5
					F	74.6172	14.3866	10
			V	**3**	**V**	74.4646	14.8462	4
					F	73.325	14.3150	12
				6	**V**	74.4646	14.8462	4
					F	73.325	14.3150	11
		V	**F**	**3**	**V**	75.0429	14.6692	3
					F	74.7321	14.6692	6
				6	**V**	74.975	14.6484	4
					F	74.685	14.4940	9
0.40	**2**	**F**	**F**	**3**	**V**	74.0144	15.2003	14
					F	69.8254	15.7239	59
				6	**V**	74.0144	15.2003	16
					F	69.8254	15.7239	57
			V	**3**	**V**	72.9351	15.4923	15
					F	66.8908	16.5263	52
				6	**V**	72.9351	15.4923	15
					F	66.8908	16.5263	50
		V	**F**	**3**	**V**	74.7245	14.7483	10
					F	73.6867	14.3508	35
				6	**V**	74.5588	14.9686	11
					F	73.1932	14.3113	44
	4	**F**	**F**	**3**	**V**	74.3101	15.0327	10
					F	71.2173	15.1287	34
				6	**V**	74.3101	15.0327	11
					F	71.2173	15.1287	39
			V	**3**	**V**	73.3439	15.3209	11
					F	68.134	16.0799	39
				6	**V**	73.3439	15.3209	9
					F	68.134	16.0799	33

Fator de Confiança	*Nº Mín Objetos*	*Reduced Error Pruning*	*Unpruned*	*Nº Folds*	*Correção MDL*	*Accuracy (%)*	*FN (%)*	*Tempo*
0.40	**4**	**V**	**F**	**3**	**V**	74.8357	14.7238	6
					F	73.8355	14.5317	24
				6	**V**	74.7396	14.7822	9
					F	73.6245	14.3320	28
	6	**F**	**F**	**3**	**V**	74.474	15.0628	9
					F	72.3211	14.8481	29
				6	**V**	74.474	15.0628	9
					F	72.3211	14.8481	27
			V	**3**	**V**	73.7281	15.1589	8
					F	69.7161	15.4923	23
				6	**V**	73.7281	15.1589	7
					F	69.7161	15.4923	26
		V	**F**	**3**	**V**	74.8507	14.8161	6
					F	74.1802	14.6579	16
				6	**V**	74.7038	14.7426	7
					F	73.8788	14.3132	21
	10	**F**	**F**	**3**	**V**	74.5211	14.9479	6
					F	73.2704	14.5580	18
				6	**V**	74.5211	14.9479	6
					F	73.2704	14.5580	18
			V	**3**	**V**	74.0107	15.0195	5
					F	71.6732	14.8368	17
				6	**V**	74.0107	15.0195	6
					F	71.6732	14.8368	16
		V	**F**	**3**	**V**	74.9336	14.7803	4
					F	74.5155	14.7275	12
				6	**V**	74.8187	14.6955	4
					F	74.1576	14.2811	12
	20	**F**	**F**	**3**	**V**	74.7264	14.8971	4
					F	74.2367	14.3056	10
				6	**V**	74.7264	14.8971	5
					F	74.2367	14.3056	11
			V	**3**	**V**	74.4646	14.8462	4
					F	73.325	14.3150	9
				6	**V**	74.4646	14.8462	5
					F	73.325	14.3150	10
		V	**F**	**3**	**V**	75.0429	14.6692	2
					F	74.7321	14.6692	7
				6	**V**	74.975	14.6484	4
					F	74.685	14.4940	10

Random Forest:

Nº Árvores	*Profund. Máx*	*Nº Atributos*	*BreakTies*	*Accuracy*	*FN (%)*	*Tempo*
20	**0**	**0**	**V**	72.1478	13.7311	16
			F	71.9463	13.7481	18
		5	**V**	71.9651	13.7989	21
			F	71.8201	13.9233	19
		10	**V**	71.3285	14.1738	33
			F	71.4377	14.0043	34
	10	**0**	**V**	75.0862	15.1890	10
			F	75.169	15.1363	10
		5	**V**	75.0165	15.2173	12
			F	75.0391	15.2154	12
		10	**V**	74.8564	15.3246	22
			F	74.8809	15.2926	19
	25	**0**	**V**	72.5528	14.2020	17
			F	72.6658	14.2567	16
		5	**V**	72.4718	14.3000	19
			F	72.2307	14.4017	19
		10	**V**	71.8973	14.4412	33
			F	71.9105	14.3716	33
50	**0**	**0**	**V**	72.6281	14.1305	44
			F	72.4398	14.2454	45
		5	**V**	72.4454	14.2171	50
			F	72.3475	14.2096	52
		10	**V**	71.9745	14.4262	1:18
			F	71.8898	14.4243	1:20
	10	**0**	**V**	75.137	15.1532	24
			F	75.1502	15.1363	24
		5	**V**	75.0579	15.1306	28
			F	75.0598	15.1532	30
		10	**V**	74.8545	15.2870	53
			F	74.8545	15.3001	56
	25	**0**	**V**	73.0086	14.4582	40
			F	72.9992	14.4149	42
		5	**V**	72.9766	14.5072	49
			F	72.7826	14.5411	50
		10	**V**	72.4191	14.5919	1:09
			F	72.3042	14.6541	1:11

Nº Árvores	*Profund. Máx*	*Nº Atributos*	*BreakTies*	*Accuracy (%)*	*FN (%)*	*Tempo*
80	**0**	**0**	**V**	72.7072	14.2246	1:03
			F	72.5961	14.3000	1:01
		5	**V**	72.6432	14.3094	1:18
			F	72.4944	14.3414	1:15
		10	**V**	72.0649	14.5147	1:59
			F	72.0819	14.4394	1:40
	10	**0**	**V**	75.1483	15.1476	30
			F	75.1992	15.1137	30
		5	**V**	75.0937	15.1080	36
			F	75.1144	15.1156	39
		10	**V**	74.8357	15.2870	1:02
			F	74.8093	15.2945	1:11
	25	**0**	**V**	73.0613	14.4846	57
			F	73.0293	14.5128	58
		5	**V**	72.9502	14.5976	1:04
			F	72.9446	14.5543	1:00
		10	**V**	72.4643	14.6484	1:51
			F	72.468	14.5900	1:46
100	**0**	**0**	**V**	72.7035	14.2811	1:21
			F	72.7298	14.3075	1:58
		5	**V**	72.5471	14.3829	2:00
			F	72.5754	14.3414	1:42
		10	**V**	72.2043	14.5241	2:48
			F	72.1591	14.4450	2:20
	10	**0**	**V**	75.1333	15.1664	40
			F	75.1766	15.1344	37
		5	**V**	75.1144	15.1174	49
			F	75.1144	15.1212	47
		10	**V**	74.8357	15.3039	1:13
			F	74.8545	15.2888	1:20
	25	**0**	**V**	73.1555	14.4751	1:20
			F	73.1367	14.4676	1:10
		5	**V**	73.018	14.5731	1:26
			F	73.0143	14.5373	1:21
		10	**V**	72.5264	14.6579	2:05
			F	72.5377	14.5995	2:08

Naive Bayes:

Kernel	*Discretização Supervisionada*	*Accuracy (%)*	*FN (%)*	*Tempo*
V	**F**	73.4249	17.3965	0
F	**V**	73.4833	17.0462	0
	F	72.566	18.2762	0

KNN:

k	*CrossValidate*	*Distância Ponder.*	*Algoritmo Pesquisa VMP*	*Accuracy (%)*	*FN (%)*	*Tempo*
10	**V**	**Não**	***LinearNNSearch***	72.453	14.9686	12:03
			BallTree	72.453	14.9686	3:50
			FilteredNSearch	65.6966	17.1950	8:53
			KDTree	72.453	14.9686	51
		1/dist	***LinearNNSearch***	72.2947	15.2154	14:07
			BallTree	72.2947	15.2154	4:08
			FilteredNSearch	65.6966	17.1950	8:59
			KDTree	72.2947	15.2154	52
		1-dist	***LinearNNSearch***	72.5396	15.3397	13:11
			BallTree	72.5396	15.3397	3:58
			FilteredNSearch	65.6966	17.1950	8:56
			KDTree	72.5396	15.3397	53
	F	**Não**	***LinearNNSearch***	72.4906	13.2019	1:32
			BallTree	72.4906	13.2019	25
			FilteredNSearch	72.4906	13.2019	54
			KDTree	72.4906	13.2019	5
		1/dist	***LinearNNSearch***	72.2947	15.2154	1:27
			BallTree	72.2947	15.2154	26
			FilteredNSearch	72.2947	15.2154	59
			KDTree	72.2947	15.2154	5
		1-dist	***LinearNNSearch***	72.5415	15.3077	1:26
			BallTree	72.5415	15.3077	23
			FilteredNSearch	72.5415	15.3077	1:05
			KDTree	72.5415	15.3077	6
20	**V**	**Não**	***LinearNNSearch***	73.5002	14.6070	15:25
			BallTree	73.5002	14.6070	4:42
			FilteredNSearch	65.6966	17.1950	9:40
			KDTree	73.5002	14.6070	1:14
		1/dist	***LinearNNSearch***	73.3288	15.0911	13:29
			BallTree	73.3288	15.0911	5:04
			FilteredNSearch	65.6966	17.1950	9:53
			KDTree	73.3288	15.0911	1:16
		1-dist	***LinearNNSearch***	73.5661	15.1137	15:02
			BallTree	73.5661	15.1137	4:57
			FilteredNSearch	65.6966	17.1950	9:25
			KDTree	73.5661	15.1137	1:13
	F	**Não**	***LinearNNSearch***	73.4833	14.0664	1:31
			BallTree	73.4833	14.0664	27
			FilteredNSearch	73.4833	14.0664	55
			KDTree	73.4833	14.0664	8
		1/dist	***LinearNNSearch***	73.3213	15.1043	1:33
			BallTree	73.3213	15.1043	27
			FilteredNSearch	73.3213	15.1043	58
			KDTree	73.3213	15.1043	8
		1-dist	***LinearNNSearch***	73.5699	15.1080	1:35
			BallTree	73.5699	15.1080	30
			FilteredNSearch	73.5699	15.1080	1:01
			KDTree	73.5699	15.1080	8

k	*CrossValidate*	*Distância Ponder.*	*Algoritmo Pesquisa VMP*	*Accuracy (%)*	*FN (%)*	*Tempo*
50	**V**	**Não**	***LinearNNSearch***	73.7714	14.7407	17:34
			BallTree	73.7714	14.7407	6:14
			FilteredNSearch	65.6966	17.1950	10:21
			KDTree	73.7714	14.7370	1:58
		1/dist	***LinearNNSearch***	73.6603	15.2455	17:04
			BallTree	73.6603	15.2455	6:20
			FilteredNSearch	65.6966	17.1950	10:31
			KDTree	73.6641	15.2455	2:01
		1-dist	***LinearNNSearch***	73.7696	15.2738	17:32
			BallTree	73.7696	15.2738	6:10
			FilteredNSearch	65.6966	17.1950	11:57
			KDTree	73.7677	15.2757	2:11
	F	**Não**	***LinearNNSearch***	73.6547	14.9517	2:05
			BallTree	73.6547	14.9517	39
			FilteredNSearch	73.6547	14.9517	1:11
			KDTree	73.6528	14.9517	12
		1/dist	***LinearNNSearch***	73.7488	15.2625	2:04
			BallTree	73.7488	15.2625	37
			FilteredNSearch	73.7488	15.2625	1:07
			KDTree	73.7413	15.2662	12
		1-dist	***LinearNNSearch***	73.7093	15.3943	1:58
			BallTree	73.7093	15.3943	40
			FilteredNSearch	73.7093	15.3943	1:04
			KDTree	73.7074	15.3943	12

MultiLayer Perceptron:

Camadas Escond.	*Tempo Treino*	*Nominal para Bin*	*Decay*	*LR*	*Momentum*	*Acuraccy (%)*	*FN (%)*	*Tempo*
a	**100**	**V**	**V**	**0.1**	**0.05**	74.393	15.0948	51
					0.1	74.4024	15.1024	50
					0.2	74.393	15.1061	49
				0.2	**0.05**	74.5155	15.1589	49
					0.1	74.5173	15.1834	48
					0.2	74.572	15.2003	48
				0.3	**0.05**	74.5211	15.2399	49
					0.1	74.5079	15.2003	49
					0.2	74.5475	15.1702	48
			F	**0.1**	**0.05**	74.6002	14.8839	50
					0.1	74.5004	14.8726	48
					0.2	74.5192	14.8255	47
				0.2	**0.05**	74.3685	14.5505	46
					0.1	74.4137	14.5185	46
					0.2	74.4533	14.5260	46
				0.3	**0.05**	74.2932	14.1135	46
					0.1	74.2668	14.2472	45
					0.2	74.2197	14.1210	46
		F	**V**	**0.1**	**0.05**	74.361	15.1476	35
					0.1	74.4081	15.1382	36
					0.2	74.4778	15.0798	36
				0.2	**0.05**	74.587	15.1382	36
					0.1	74.5908	15.1721	36
					0.2	74.6059	15.1551	36
				0.3	**0.05**	74.636	15.0986	36
					0.1	74.6304	15.0948	36
					0.2	74.7321	14.9649	36
			F	**0.1**	**0.05**	74.8319	14.3000	36
					0.1	74.8225	14.3923	36
					0.2	74.7528	14.4168	36
				0.2	**0.05**	74.3968	14.4638	36
					0.1	74.4721	14.2378	36
					0.2	74.2894	14.1775	36
				0.3	**0.05**	74.3535	14.2887	36
					0.1	74.3441	14.3979	36
					0.2	74.2273	14.6428	36
	300	**V**	**V**	**0.1**	**0.05**	74.3968	15.1061	2:26
					0.1	74.4156	15.0930	2:21
					0.2	74.4232	15.1306	2:24
				0.2	**0.05**	74.5795	15.1325	2:22
					0.1	74.5738	15.1495	2:21
					0.2	74.5324	15.1400	2:16
				0.3	**0.05**	74.5983	15.1382	2:18
					0.1	74.6115	15.1193	2:16
					0.2	74.6285	15.1005	2:16
			F	**0.1**	**0.05**	74.6059	14.9234	2:18
					0.1	74.5362	14.8029	2:18

Camadas Escond.	*Tempo Treino*	*Nominal para Bin*	*Decay*	*LR*	*Momentum*	*Acuraccy (%)*	*FN (%)*	*Tempo*
a	**300**	**V**	**F**	**0.1**	**0.2**	74.5795	14.7633	2:20
				0.2	**0.05**	74.3949	14.5919	2:16
					0.1	74.3403	14.5486	2:17
					0.2	74.3761	14.6823	2:21
				0.3	**0.05**	74.2273	14.0777	2:18
					0.1	74.2988	14.0363	2:19
					0.2	74.2254	14.0495	2:21
		F	**V**	**0.1**	**0.05**	74.4816	15.0760	1:55
					0.1	74.4985	15.0779	1:53
					0.2	74.5362	15.0609	1:50
				0.2	**0.05**	74.6228	15.0835	1:55
					0.1	74.6002	15.1061	1:56
					0.2	74.6756	15.0704	1:50
				0.3	**0.05**	74.6548	15.0289	1:54
					0.1	74.6944	15.0195	1:54
					0.2	74.8771	14.8971	1:53
			F	**0.1**	**0.05**	74.7867	14.3282	1:53
					0.1	74.847	14.4281	1:51
					0.2	74.7754	14.4733	1:54
				0.2	**0.05**	74.4326	14.4337	1:50
					0.1	74.4533	14.2623	1:53
					0.2	74.2951	14.2793	1:49
				0.3	**0.05**	74.2988	14.3621	1:52
					0.1	74.3798	14.3960	1:48
					0.2	74.1971	14.6239	1:52
o	**100**	**V**	**V**	**0.1**	**0.05**	74.2555	14.8745	18
					0.1	74.2631	14.8801	18
					0.2	74.3007	14.8217	18
				0.2	**0.05**	74.2913	14.7445	18
					0.1	74.2913	14.7445	18
					0.2	74.2838	14.7407	18
				0.3	**0.05**	74.312	14.7181	18
					0.1	74.297	14.7464	18
					0.2	74.3403	14.7200	19
			F	**0.1**	**0.05**	73.6113	14.9178	18
					0.1	73.5925	14.9686	18
					0.2	73.4701	15.1005	18
				0.2	**0.05**	73.1122	15.3529	18
					0.1	73.0482	15.3359	18
					0.2	72.9389	15.3058	19
				0.3	**0.05**	72.8203	14.8933	18
					0.1	72.7035	14.8594	18
					0.2	72.6093	14.6315	18
		F	**V**	**0.1**	**0.05**	74.1858	14.9743	16
					0.1	74.1802	14.9762	17
					0.2	74.1934	14.9724	16
				0.2	**0.05**	74.1726	14.9479	16
					0.1	74.1651	14.9479	17
					0.2	74.1915	14.9178	16

Camadas Escond.	*Tempo Treino*	*Nominal para Bin*	*Decay*	*LR*	*Momentum*	*Acuraccy (%)*	*FN (%)*	*Tempo*
o	**100**	**F**	**V**	**0.3**	**0.05**	74.2536	14.8349	16
					0.1	74.2273	14.8368	16
					0.2	74.2103	14.8387	17
			F	**0.1**	**0.05**	73.6302	14.8330	17
					0.1	73.6245	14.8519	16
					0.2	73.5906	14.8820	17
				0.2	**0.05**	73.2365	15.2493	17
					0.1	73.2177	15.2719	17
					0.2	73.1273	15.2832	16
				0.3	**0.05**	72.8899	15.0986	17
					0.1	72.8466	15.0722	16
					0.2	72.7939	15.0251	17
	300	**V**	**V**	**0.1**	**0.05**	74.2838	14.8575	58
					0.1	74.3007	14.8650	55
					0.2	74.3196	14.8236	57
				0.2	**0.05**	74.3629	14.8198	58
					0.1	74.3516	14.8311	55
					0.2	74.3648	14.8330	55
				0.3	**0.05**	74.3497	14.8745	54
					0.1	74.3196	14.9027	55
					0.2	74.3459	14.8707	55
			F	**0.1**	**0.05**	73.6358	14.8707	56
					0.1	73.6	14.9121	56
					0.2	73.4964	15.0496	55
				0.2	**0.05**	73.116	15.3604	55
					0.1	73.0689	15.3322	55
					0.2	72.9427	15.3039	55
				0.3	**0.05**	72.824	14.8989	55
					0.1	72.6959	14.8632	54
					0.2	72.6093	14.6390	54
		F	**V**	**0.1**	**0.05**	74.216	14.9479	52
					0.1	74.2273	14.9479	50
					0.2	74.2574	14.9366	51
				0.2	**0.05**	74.2442	14.9592	53
					0.1	74.2574	14.9536	51
					0.2	74.28	14.9611	52
				0.3	**0.05**	74.2744	14.9743	51
					0.1	74.28	14.9818	51
					0.2	74.3045	14.9743	51
			F	**0.1**	**0.05**	73.6302	14.7972	51
					0.1	73.6377	14.8104	50
					0.2	73.5869	14.8688	51
				0.2	**0.05**	73.2403	15.2474	54
					0.1	73.2196	15.2700	52
					0.2	73.1235	15.2813	52
				0.3	**0.05**	72.8862	15.1005	50
					0.1	72.8485	15.0722	53
					0.2	72.7958	15.0233	49

ANEXO IV – MODELOS GERADOS PELAS TÉCNICAS RF, NB E MLP (CENÁRIO I)

Random Forest:

Text

```
|   |   |   |   |   |   |   |   |   |   |   Idade >= 62 : Doença (1/0)
|   |   |   |   |   |   |   |   |   |   |   AntecedentesFam = Sim : Doença (2/0)
|   |   |   |   |   |   |   |   |   |   Colesterol_Total >= 176.5 : Doença (11/0)

Size of the tree : 19239
Max depth of tree: 25


RandomTree
==========

PA_Alta < 129.5
|   Sexo = Feminino
|   |   Colesterol_Total < 243.5
|   |   |   Idade < 52.5
|   |   |   |   Idade < 43.5
|   |   |   |   |   IMC < 23.85
|   |   |   |   |   |   PA_Baixa < 79.5
|   |   |   |   |   |   |   Colesterol_Total < 182.5
|   |   |   |   |   |   |   |   AntecedentesFam = Não
|   |   |   |   |   |   |   |   |   Exercicio_Fisico = Baixo : Sem Doença (37/0)
```

Naive Bayes:

Text

```
Naive Bayes Classifier

                          Class
Attribute                Doença Sem Doença
                         (0.47)     (0.53)
=============================================
Idade
  '(-inf-40.5]'           580.0     2153.0
  '(40.5-44.5]'          1680.0     3710.0
  '(44.5-52.5]'          6590.0     9599.0
  '(52.5-54.5]'          2691.0     3057.0
  '(54.5-58.5]'          5705.0     5168.0
  '(58.5-60.5]'          2924.0     2137.0
  '(60.5-inf)'           4823.0     2288.0
  [total]               24993.0    28112.0

Sexo
  Feminino              11755.0    19327.0
  Masculino             13233.0     8780.0
  [total]               24988.0    28107.0
```

MultiLayer Perceptron:

Text

```
    Node 10    -0.6660704274317464
Sigmoid Node 1
    Inputs    Weights
    Threshold    -0.12161588060611245
    Node 2    -0.8879429521630176
    Node 3    -1.5072708007954967
    Node 4    1.3683486970794663
    Node 5    0.7504647518707915
    Node 6    -0.7049081856568135
    Node 7    -1.6176110265892418
    Node 8    -2.5595536069119524
    Node 9    0.7814948571362955
    Node 10    0.6660704274317467
Sigmoid Node 2
    Inputs    Weights
    Threshold    -2.573409931623208
    Attrib Idade    3.2897348691265886
    Attrib Sexo=Masculino    6.711749218243419
    Attrib Dor_Esforco=Sim    1.1645428631578 79
    Attrib PA_Alta    -2.0192311399379386
    Attrib PA_Baixa    -3.088070801885909
```

ANEXO V – CARACTERIZAÇÃO DAS TABELAS DE DIMENSÃO DO DW

Dimensão Calendário:

Caracterização	**Identificação**	DimCalendário				
	Descrição	Caracterização das principais informações cronológicas				
	Tipo	Dimensão sem variação				
	Dimensão	10 MB				
	Crescimento	Não tem. O povoamento é efetuado para um horizonte temporal de 20 anos desde a data de início do DW.				
Atributos	**Identificação**	**Descrição**	**Chave**	**Domínio**	**Variação**	**Exemplos**
	Data	Data do calendário	PK	Data	Não	25/08/2019
	Semana	Nome do dia da semana	Não	*String*	Não	Domingo
	Mês	Número do mês	Não	Inteiro	Não	8
	Trimestre	Número do trimestre	Não	Inteiro	Não	3
	Ano	Número do ano	Não	Inteiro	Sim	2019
Hierarquias	**Número**	**Identificação**	**Esquema**			
	1	H1	Data → Mês → Trimestre → Ano → *ALL*			
	2	H2	Data → Semana → *ALL*			
Observações						

Dimensão Distrito:

Caracterização	**Identificação**	DimDistrito
	Descrição	Caracterização dos distritos de Portugal, com informações relativas à província, região e zona costeira em que se inserem.
	Tipo	Dimensão sem variação
	Dimensão	1 MB
	Crescimento	Não tem. Apenas se efetua um carregamento de dados inicial.

Atributos	**Identificação**	**Descrição**	**Chave**	**Domínio**	**Variação**	**Exemplos**
	DistritoId	Código de três letras para a identificação do distrito	PK	*String*	Não	BRG
	Designação	Nome do distrito	Não	*String*	Não	Braga
	Província	Nome da província do distrito	Não	*String*	Não	Minho
	Região	Nome da região do distrito	Não	*String*	Não	Norte
	Localização Costeira	Descrição da localização do distrito	Não	*String*	Não	Litoral

Hierarquias	**Número**	**Identificação**	**Esquema**
	1	H1	DistritoID → Província → Região → *ALL*
	2	H2	DistritoID → Localização Costeira → *ALL*

Observações
À partida, não se devem verificar alterações nos distritos de Portugal no período de vida do DW. Contudo, caso ocorram, é preciso eliminar e povoar de novo esta dimensão, com dados atuais.

Dimensão Utilizador:

Caracterização	**Identificação**	DimUtilizador
	Descrição	Caracterização das principais informações pessoais relativas a cada um dos utilizadores.
	Tipo	Dimensão com variação
	Dimensão	10 MB
	Crescimento	5%/Ano

	Identificação	**Descrição**	**Chave**	**Domínio**	**Variação**	**Exemplos**
Atributos	UtilizadorId	Código identificativo do utilizador	PK	Inteiro	Não	1111
	Sexo	Género do utilizador	Não	*String*	Não	Feminino
	Local	Local de residência do utilizador	Não	*String*	Sim	Braga
	Habilitações	Habilitações literárias do utilizador	Não	*String*	Sim	Mestrado
	Rendimento	Rendimento anual líquido do utilizador	Não	*String*	Sim	Elevado

	Número	**Identificação**	**Esquema**
Hierarquias	1	H1	Utilizador → Sexo → *ALL*
	2	H2	Utilizador → Local → *ALL*
	3	H3	Utilizador → Habilitações → *ALL*
	4	H4	Utilizador → Rendimento → *ALL*

Observações

ANEXO VI – CARACTERIZAÇÃO DA TABELA DE FACTOS DO DW

Caracterização	**Identificação**	TFBemEstar			
	Descrição	Tabela que reúne todos os índices de bem-estar cardíaco calculados pelos utilizadores			
	Tipo	Transacional			
	Utilidade Estratégica	Incentivar a adoção de um estilo de vida saudável. Detetar precocemente possíveis casos de DCV. Reduzir a incidência de casos de DCV.			
	Povoamento	Diário, entre as 00:00 e as 01:00.			
	Dimensão	10 MB			
	Crescimento	5%/Mês			
Atributos	**Identificação**	**Descrição**	**Chave**	**Domínio**	**Exemplos**
	Data	Data do calendário	FK	Data	25/08/2019
	Semana	Nome do dia da semana	Não	*String*	Domingo
	Mês	Número do mês	Não	Inteiro	8
	Trimestre	Número do trimestre	Não	Inteiro	3
	Ano	Número do ano	Não	Inteiro	2019
Medidas	**Identificação**	**Descrição**	**Tipo**	**Domínio**	**Exemplos**
	Idade	Idade do utilizador	Não Agr.	Inteiro	56
	Sexo	Género do utilizador	Não Agr.	*String*	Feminino
	IMC	IMC do utilizador	Não Agr.	*Float*	25.15
	Antecedentes	Historial familiar de DCV	Não Agr.	*String*	Não
	QtddCigarros	Quantidade de cigarros fumados	Não Agr.	*Float*	233.60
	ColesterolTotal	Nível de colesterol total	Não Agr.	Inteiro	150
	Diabetes	Identificação de indivíduo diabético	Não Agr.	*String*	Não
	Hipertensão	Identificação de indivíduo hipertenso	Não Agr.	*String*	Não
	DorEsforço	Identificação de dor após esforço	Não Agr.	*String*	Não
	Hipotiroidismo	Identificação de hipotiroidismo	Não Agr.	*String*	Não
	PA Alta	Nível de pressão arterial alta	Não Agr.	Inteiro	120
	PA Baixa	Nível de pressão arterial baixa	Não Agr.	Inteiro	85
	ExercícioFísico	Nível de exercício físico praticado	Não Agr.	*String*	Moderado
	Índice	Índice de bem-estar cardíaco não ponderado	Agr.*	*Float*	1.92
Observações					
* Agregável por média e por média ao longo do tempo.					

ANEXO VII – PERFIS DE DADOS

Fonte 1:

Perfis-chave de Candidato - [dbo].[Fonte1]

Colunas de Chaves	Intensidade de Chave
id	100.0000 %

Perfis de Distribuição de Comprimento de Coluna - [dbo].[Fonte1]

Coluna	Comprimento Mínimo	Comprimento Máximo	Ignorar Espaços à Esquerda	Ignorar Espaços à Direita
antecedentes	3	3	☐	☑
diabetes	3	3	☐	☑
distrito	3	3	☐	☑
dor_esforco	3	3	☐	☑
fumador	3	3	☐	☑
habilitacoes	6	12	☐	☑
hipertensao	3	3	☐	☑

Distribuição de Comprimento - habilitacoes — Conexão Criptografada — 1000 Linhas

Comprimento	Contar	Porcentagem
12	2004	39.7461 %
6	1036	20.5474 %
10	1045	20.7259 %
8	957	18.9806 %

Perfis de Distribuição de Valor de Coluna - [dbo].[Fonte1]

Coluna	Número de Valores Distintos
sexo	2
rendimento	71
peso	123
num_cigarros	5
num_anos_fum	36
id	5042
hipotiroidismo	2
hipertensao	2
habilitacoes	5
glicose_rapida	106
fumador	2
dor_esforco	2
diabetes	2
data_nascimento	1214
data_analises	93
concelho	20
colesterol_total	152
antecedentes	2
altura	48

Perfis de Razão Nula de Coluna - [dbo].[Fonte1] — Conexão Criptografada — 1000 Linhas

Coluna	Contagem Nula	Porcentagem Nula
altura	0	0.0000 %
antecedentes	0	0.0000 %
colesterol_total	0	0.0000 %
data_analises	0	0.0000 %
data_nascimento	0	0.0000 %
diabetes	0	0.0000 %
distrito	0	0.0000 %
dor_esforco	0	0.0000 %
fumador	0	0.0000 %
glicose_rapida	0	0.0000 %
habilitacoes	0	0.0000 %
hipertensao	0	0.0000 %
hipotiroidismo	0	0.0000 %
id	0	0.0000 %
num_anos_fum	0	0.0000 %
num_cigarros	0	0.0000 %
peso	0	0.0000 %
rendimento	0	0.0000 %
sexo	0	0.0000 %

Fonte 2:

Perfis de Distribuição de Valor de Coluna - [dbo].[Fonte2]

Coluna	Número de Valores Distintos
calorias	2252
data_índice	765
data_registo	774
Id	5044
pa_alta	68
pa_baixa	52

Perfis de Inclusão - [dbo].[Fonte2]

Colunas Laterais do Subconjunto	Colunas Laterais do Superconjunto	Intensidade de Inclusão
[dbo].[Fonte2]([Id])	[dbo].[Fonte1]([id])	99.9980 %

Violações de Inclusão — Conexão Criptografada — 1000 Linhas

Id	Contar	Porcentagem
7000	1	0.0010 %
6000	1	0.0010 %

Perfis de Razão Nula de Coluna - [dbo].[Fonte2] — Conexão Criptografada — 1000 Linhas

Coluna	Contagem Nula	Porcentagem Nula
calorias	0	0.0000 %
data_índice	0	0.0000 %
data_registo	0	0.0000 %
Id	0	0.0000 %
pa_alta	0	0.0000 %
pa_baixa	0	0.0000 %

ANEXO VIII – *TRIGGERS* E *STORED PROCEDURES*

***Trigger* de Inserção:**

```
CREATE TRIGGER [dbo].[after_insert_fonte1]
       ON [dbo].[Fonte1]
       AFTER INSERT
AS
BEGIN
       SET NOCOUNT ON;
       DECLARE @id int
       SELECT @id = id FROM inserted
       EXEC [dbo].[NovoUtilizador] @id
END
```

***Trigger* de Modificação:**

```
CREATE TRIGGER [dbo].[after_update_fonte1]
   ON  [dbo].[Fonte1]
   AFTER UPDATE
AS
IF UPDATE(habilitacoes) OR UPDATE(rendimento) or UPDATE(distrito)
BEGIN
       SET NOCOUNT ON;
       DECLARE @id int
       SELECT @id = id FROM inserted
       EXEC [dbo].[AtualizarUtilizador] @id
END
```

***Trigger* de Remoção:**

```
CREATE TRIGGER [dbo].[for_delete_fonte1]
       ON [dbo].[Fonte1]
       INSTEAD OF DELETE
AS
BEGIN
       SET NOCOUNT ON;
       DECLARE @id int
       SELECT @id = id FROM deleted
          EXEC [dbo].[RemoverUtilizador] @id
          DELETE FROM Fonte1 where id = @id;
END
GO
```

***Stored Procedure* de Inserção:**

```
CREATE PROCEDURE [dbo].[NovoUtilizador](@novo_utilizador int)
AS
BEGIN
DECLARE @idUt int;
DECLARE @sex VARCHAR(100);
DECLARE @hab VARCHAR(100);
DECLARE @rendim int;
DECLARE @loc VARCHAR(100);

SELECT @idUt = id, @sex = sexo, @hab = habilitacoes, @rendim = rendimento, @loc =
distrito from FontesDados.dbo.Fonte1 where FontesDados.dbo.Fonte1.id =
@novo_utilizador;

INSERT INTO DSA.dbo.audDimUtilizador (id_utilizador, sexo, habilitacoes,
rendimento, local, operacao, datahora) VALUES (@idUt, @sex, @hab, @rendim, @loc,
'novo', GETDATE());
END
GO
```

***Stored Procedure* de Atualização:**

```
CREATE PROCEDURE [dbo].[AtualizarUtilizador](@atual_utilizador int)
AS
BEGIN
DECLARE @idUt int;
DECLARE @sex VARCHAR(100);
DECLARE @hab VARCHAR(100);
DECLARE @rendim int;
DECLARE @loc VARCHAR(100);

SELECT @idUt = id, @sex = sexo, @hab = habilitacoes, @rendim = rendimento, @loc =
distrito from FontesDados.dbo.Fonte1 where FontesDados.dbo.Fonte1.id =
@atual_utilizador;

INSERT INTO DSA.dbo.audDimUtilizador (id_utilizador, sexo, habilitacoes,
rendimento, local, operacao, datahora) VALUES (@idUt, @sex, @hab, @rendim, @loc,
'atualizado', GETDATE());
END
GO
```

***Stored Procedure* de Remoção:**

```
CREATE PROCEDURE [dbo].[RemoverUtilizador](@remov_utilizador int)
AS
BEGIN
DECLARE @idUt int;
DECLARE @sex VARCHAR(100);
DECLARE @hab VARCHAR(100);
DECLARE @rendim int;
DECLARE @loc VARCHAR(100);

SELECT @idUt = id, @sex = sexo, @hab = habilitacoes, @rendim = rendimento, @loc =
distrito from FontesDados.dbo.Fonte1 where FontesDados.dbo.Fonte1.id =
@remov_utilizador;

INSERT INTO DSA.dbo.audDimUtilizador (id_utilizador, sexo, habilitacoes,
rendimento, local, operacao, datahora) VALUES (@idUt, @sex, @hab, @rendim, @loc,
'removido', GETDATE());
END
GO
```

ANEXO IX – EXTRATO DE *EMAILS* DE CONFIRMAÇÃO DE SUCESSO DE DM E ETL

***Email* de Confirmação de Sucesso de DM:**

Gmail

DM Concluído

Ana <droptese@gmail.com>
Para: droptese@gmail.com

O processo de DM foi concluído com sucesso!

Job:

JobName : MainJobDM
Directory : /
JobEntry : DM Concluído

Message date: 2019/10/19 16:14:17.014

Previous results:

Job entry Nr : 3
Errors : 0
Lines read : 0
Lines written : 0
Lines input : 0
Lines output : 0
Lines updated : 0
Lines rejected : 0
Script exist status : 0
Result : true

Path to this job entry:

MainJobDM
MainJobDM : : start : Start of job execution (2019/10/19 16:14:13.796)
MainJobDM : : Start : start : Start of job execution (2019/10/19 16:14:13.797)
MainJobDM : : Start : [nr=0, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:13.797)
MainJobDM : : Contar Linhas Dataset DM : Followed unconditional link : Start of job execution (2019/10/19 16:14:13.798)
MainJobDM : : Contar Linhas Dataset DM : [nr=1, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:13.829)
MainJobDM : : Registos >= 5000 : Followed unconditional link : Start of job execution (2019/10/19 16:14:13.829)
MainJobDM : : Registos >= 5000 : [nr=1, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:13.829)
MainJobDM : : DM : Followed unconditional link : Start of job execution (2019/10/19 16:14:13.830)
DataMining
DataMining : : Start : Start of job entry : Start of job execution (2019/10/19 16:14:13.840)
DataMining : : Start : [nr=3, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:13.841)
DataMining : : Treino dos Modelos : Followed unconditional link : Start of job execution (2019/10/19 16:14:13.842)
DataMining : : Treino dos Modelos : [nr=5, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:16.865)
DataMining : : Scores e Seleção do Melhor Modelo : Followed link after success : Start of job execution (2019/10/19 16:14:16.865)
DataMining : : Scores e Seleção do Melhor Modelo : [nr=6, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:17.009)
DataMining : : Success : Followed link after success : Start of job execution (2019/10/19 16:14:17.009)
DataMining : : Success : [nr=6, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:17.010)
MainJobDM : : DM : [nr=3, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:17.011)
MainJobDM : : Success : Followed link after success : Start of job execution (2019/10/19 16:14:17.011)
MainJobDM : : Success : [nr=3, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/19 16:14:17.011)
MainJobDM : : DM Concluído : Followed link after success : Start of job execution (2019/10/19 16:14:17.012)

***Email* de Confirmação de Sucesso de ETL:**

Gmail

ETL Concluído

Ana <droptese@gmail.com>
Para: droptese@gmail.com

O processo de ETL foi concluído com sucesso!

Job:

JobName : MainJobETL
Directory : /
JobEntry : ETL Concluído

Message date: 2019/10/16 19:59:32.818

Previous results:

Job entry Nr : 1
Errors : 0
Lines read : 0
Lines written : 0
Lines input : 0
Lines output : 0
Lines updated : 0
Lines rejected : 0
Script exist status : 0
Result : true

Path to this job entry:

MainJobETL
MainJobETL : : start : Start of job execution (2019/10/16 19:50:48.485)
MainJobETL : : Start : start : Start of job execution (2019/10/16 19:50:48.486)
MainJobETL : : Start : [nr=0, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:48.486)
MainJobETL : : ETL : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.487)
ETL
ETL : : Start : Start of job entry : Start of job execution (2019/10/16 19:50:48.496)
ETL : : Start : [nr=1, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:48.496)
ETL : : ETL1 : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.497)
ETL1
ETL1 : : Start : Start of job entry : Start of job execution (2019/10/16 19:50:48.509)
ETL1 : : Start : [nr=3, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:48.509)
ETL1 : : linhas DimDistrito = 0 : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.510)
ETL1 : : linhas DimCalendário = 0 : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.510)
ETL1 : : DimUtilizador : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.511)
ETL1 : : PreTF : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.518)
DimUtilizador
DimUtilizador : : Start : Start of job entry : Start of job execution (2019/10/16 19:50:48.523)
DimUtilizador : : Start : [nr=5, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:48.524)
DimUtilizador : : Extração Utilizador : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.525)
DimUtilizador : : Extração Utilizador : [nr=7, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:53.492)
DimUtilizador : : Limpeza Utilizador : Followed link after success : Start of job execution (2019/10/16 19:50:53.492)
DimUtilizador : : Limpeza Utilizador : [nr=8, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:58.991)
DimUtilizador : : Carregar Utilizador : Followed link after success : Start of job execution (2019/10/16 19:50:58.991)
CarregDimUtilizador
CarregDimUtilizador : : Start : Start of job entry : Start of job execution (2019/10/16 19:50:59.002)
CarregDimUtilizador : : Start : [nr=9, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:59.003)
CarregDimUtilizador : : Gerar SK : Followed unconditional link : Start of job execution (2019/10/16 19:50:59.003)
CarregDimUtilizador : : Gerar SK : [nr=11, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:59.028)
CarregDimUtilizador : : SKGUtilizador : Followed unconditional link : Start of job execution (2019/10/16 19:50:59.028)
CarregDimUtilizador : : SKGUtilizador : [nr=12, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:51:03.966)
CarregDimUtilizador : : Carregar DW : Followed unconditional link : Start of job execution (2019/10/16 19:51:03.966)
CarregDimUtilizador : : Carregar DW : [nr=13, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:51:10.398)
CarregDimUtilizador : : Success : Followed link after success : Start of job execution (2019/10/16 19:51:10.399)
CarregDimUtilizador : : Success : [nr=13, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:51:10.399)
DimUtilizador : : Carregar Utilizador : [nr=9, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:51:10.406)
DimUtilizador : : Success : Followed link after success : Start of job execution (2019/10/16 19:51:10.406)
DimUtilizador : : Success : [nr=9, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:51:10.406)
TF
TF : : Start : Start of job entry : Start of job execution (2019/10/16 19:50:48.530)
TF : : Start : [nr=5, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:48.531)
TF : : Extração TF : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.532)
ExtraçãoTF
ExtraçãoTF : : Start : Start of job entry : Start of job execution (2019/10/16 19:50:48.547)
ExtraçãoTF : : Start : [nr=7, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:50:48.547)
ExtraçãoTF : : Extrair TF2 : Followed unconditional link : Start of job execution (2019/10/16 19:50:48.548)
ExtraçãoTF : : Extrair TF2 : [nr=9, errors=0, exit_status=0, result=true] : Job execution finished (2019/10/16 19:56:50.709)
ExtraçãoTF : : Extrair TF1 : Followed link after success : Start of job execution (2019/10/16 19:56:50.709)